Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

“Only Those Who Understand Will Understand”: Collaborative Satirical Expression under Risk on Chinese Social Media

Yibo Meng

mengyb22@tsinghua.org.cn

Tsinghua University

Yuqi Zhang

Harvard Business School

Bingyi Liu

University of Michigan

Lionel Z. Wang

Hong Kong Polytechnic University

Research Article

Keywords: political expression, satire, social media, collaborative expression, content moderation, China, expressive engineering

Posted Date: July 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9647528/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

"Only Those Who Understand Will Understand": Collaborative Satirical Expression under Risk on Chinese Social Media

Yibo Meng[1*], Yuqi Zhang[2], Bingyi Liu[3], Lionel Z. Wang[4]

[1*]Tsinghua University, Beijing, China.
[2]Harvard Business School, Boston, MA, United States.
[3]University of Michigan, Ann Arbor, Ann Arbor, MI, United States.
[4]Hong Kong Polytechnic University, Hong Kong, China.

*Corresponding author(s). E-mail(s): mengyb22@tsinghua.org.cn;
Contributing authors: yzhang@hbs.edu; bingyi@umich.edu;
lionel-z.wang@connect.polyu.hk;

Abstract

In environments where political expression entails potential risks, social media users often avoid direct statements and instead participate in public discussions through humor, satire, and indirect references. However, existing research has primarily focused on platform governance or individual evasion strategies, leaving the collaborative processes through which such expressions are produced, circulated, and interpreted underexplored. This paper investigates how Chinese social media users employ satirical expression to engage with political issues under perceived risk conditions. Drawing on semi-structured interviews with 26 Chinese social media users and approximately 104 examples of satirical expressions shared by participants, we use grounded theory–informed thematic analysis combined with process tracing to examine the structural forms, motivations, and interactional dynamics of these expressions. Our findings show that political meanings are rarely conveyed through direct statements; instead, they emerge through collaborative interaction. We identify seven recurring expressive structures and demonstrate how these forms correspond to motivations such as risk avoidance, emotion regulation, social presence, and group safety. We further reveal a collaborative process—triggering, encoding, relay, and stabilization—through which satire becomes collectively interpretable while remaining deniable. We conceptualize this practice as *expressive engineering*, describing how users strategically design expressions under constraints of platform governance, audience structure, and perceived risk.

Keywords: political expression, satire, social media, collaborative expression, content moderation, China, expressive engineering

1 Introduction

In contemporary society, public political expression is regarded as an important component of political participation [1]. With the development of social media platforms, a large amount of political discussion, opinion expression, and interaction around public issues has gradually shifted from traditional public spaces to digital platforms [1, 2]. Social media not only provides the public with channels to express opinions, but also offers infrastructures for large-scale online interaction, collaborative discussion, and the aggregation of viewpoints [3, 4]. Studies have shown that digital platforms, through recommendation algorithms, interaction mechanisms, and social network structures, enable users with similar political positions to form clusters, reinforce their views, and in some cases produce clear phenomena of group polarization [5–7]. Therefore, social media has become an important arena for understanding contemporary political expression and public discussion.

However, social media is not a completely open space for expression. Platform governance, content moderation, and broader socio-political environments all exert profound influences on online expression. In research including the HCI and CSCW communities, content moderation is widely regarded as an important component of platform governance [8–13]. Through establishing rules, identifying violating content, and implementing actions such as removal or account suspension, moderation shapes who can participate, what content can be seen, and which forms of speech are suppressed [14, 15]. Existing research has pointed out that platform moderation mechanisms are often highly opaque and uncertain, making it difficult for users to accurately understand the specific reasons why content is removed or restricted [10]. Under such circumstances, users often develop "folk theories" about platform rules and algorithmic mechanisms through continuous experimentation and accumulated experience, and adjust their expression styles and behavioral strategies accordingly [16, 17].

When political issues enter social media discussions, this expressive environment often becomes more complex. In certain countries and regions, political content may be subject to stricter regulation and censorship, making public expression potentially risky [18, 19]. Existing research shows that in such environments, users often develop various adaptive strategies, including self-censorship, lexical variation, metaphorical expression, and humorous satire, in order to reduce expressive risks while maintaining the possibility of participating in public discussions [20–23]. These strategies not only demonstrate how individuals adapt their expression under constrained environments, but also reflect how users negotiate their positions between technological systems and social structures.

In practice, political expression does not always appear in the form of direct statements or explicit political positions [24]. Particularly in environments where expression carries potential risks, people often discuss political topics through joking, satire, metaphor, or imitation, thereby engaging in a form of "deconstructive" discussion of

politics [23, 25, 26]. These expressions often appear on the surface as humor or everyday jokes, yet carry clear political meanings within specific contexts. For example, irony may be conveyed through excessive praise, reality may be referenced through historical metaphors, or satire may be produced through imitation of official discourse. These forms of expression can be quickly understood within a group, while simultaneously maintaining a certain degree of ambiguity for external observers, thereby forming a mode of expression characterized by deniability.

More importantly, such expressions are often not produced independently by a single user, but gradually emerge through collaborative interaction [2, 25]. For example, a sarcastic comment may be repeatedly echoed, modified, or reinforced within the comment section, gradually forming a template of expression with relatively stable meaning through group interaction. In this process, users not only co-produce linguistic forms but also jointly maintain interpretive rules, such as the boundary implied by the phrase "those who understand will understand." This suggests that in constrained expressive environments, the production of political meaning is not merely an individual behavior but rather a collaborative process of meaning construction.

Consider the phrase "They really won again" (真是又赢了) appearing in a comment section under a piece of news about an economic setback. To an outsider or an automated algorithm, this statement appears to be a positive affirmation of national strength, aligning perfectly with mainstream patriotic narratives. However, to the community of users engaged in the discussion, the meaning is instantly inverted: it is a biting satire, mocking the disconnect between official triumphalism and the lived reality of the public. This phenomenon, colloquially known as "Those who understand will understand" (懂的都懂), represents a sophisticated form of communication where meaning is detached from literal text and relocated into the shared, tacit context of the community.

This vignette illustrates the critical challenge facing political expression in such environments: the boundary of permissible speech is a "moving target," fluctuating with the political sensitivity of the moment [27, 28]. Consequently, users cannot rely on fixed rules to guide their speech. Instead, they must engage in a continuous, high-stakes game of cat-and-mouse, where direct expression carries the risk of account suspension or social retaliation, while complete silence equates to withdrawal from public discourse.

However, while this "cat-and-mouse" dynamic is well-recognized, a significant gap remains in HCI and CSCW research regarding how such expressions are *collaboratively* constructed and maintained. Existing research has revealed various ways in which users respond to content moderation and censorship environments [10, 14], but most work still focuses primarily on two levels. The first line of research focuses on governance mechanisms at the platform level, such as moderation rules, algorithmic decision-making, and the influence of platform policies on user behavior [8]. The second line of research focuses on expression strategies at the individual level, such as self-censorship, lexical substitution, or technical methods used to evade algorithmic detection [18, 19, 22]. Although phenomena such as political humor, online satire, and "e'gao culture" have received certain attention in communication and media studies [23, 25, 26], research on these forms of expression remains relatively limited within

CSCW and HCI. Fewer studies have examined how users *collaboratively* construct and maintain a form of political *expression* that is understandable yet intentionally ambiguous through interactive processes among users. In particular, several questions still lack systematic understanding: Why do people choose to discuss political issues through joking and satire when expressive risks exist? How are such expressions encoded, disseminated, and collectively understood through social media interaction? And how do platform mechanisms, audience structures, and risk perceptions shape the specific structure of these expressions? What are the consequences caused by such expression?

Therefore, this paper explores how young Chinese social media users discuss political issues through joking, satire, and deconstructive expressions on social media by conducting semi-structured interviews with 26 participants. Rather than simply treating humor or satire as rhetorical strategies, this study focuses on how these expressions are produced, circulated, and gradually form shared understanding through interaction.

Based on the analysis of interview materials and expression cases, we observe that in constrained expressive environments, political meanings are often not formed through direct statements but are constructed and maintained through a series of gradually unfolding interactive processes. For example, some expressions initially appear in ambiguous or deniable forms, and are subsequently repeated, modified, or supplemented within comment interactions, gradually forming expressions recognizable within a group. This process indicates that political expression in such environments is not merely an individual behavior but also a meaning-construction practice that relies on collaborative interaction.

On this basis, this paper attempts to propose an analytical perspective for understanding this phenomenon: these expressive practices can be understood as a form of "expressive engineering" conducted under multiple constraints. In this process, expressions need to simultaneously satisfy several conditions, such as being understandable to specific audiences, being able to circulate through interaction, while also maintaining a certain level of ambiguity or deniability. Through this perspective, we can better understand how users continuously adjust their expressive practices among platform rules, audience structures, and dynamic risk expectations.

The contributions of this paper are mainly threefold. First, through empirical research, this study illustrates a collaborative expressive process that gradually emerges through interaction in constrained expressive environments, thereby complementing existing research that primarily focuses on individual expression strategies. Second, this paper proposes the conceptual lens of "expressive engineering" to describe how users design expressive forms under multiple constraints. Third, this paper further discusses how platform distribution mechanisms, audience structures, and risk expectations jointly shape these expressive forms, providing new empirical materials and analytical perspectives for understanding political expression on social media.

2 Related Works

2.1 Collaborative Expression in Limited Political Context

Public political expression is an indispensable part of political participation in the modern society, and social media complicates such participation by providing platforms for online massive engagement - social media platforms support online political activism, raising collective participation and coordination [29–31], through which people with similar ideologies are gathered by algorithms to create homophily and even polarization [32–34].

However, social media are not anarchic, but are instead governed by opaque, often biased content moderation. Content moderation, including practices of creating guidelines and rules, detecting violated content, taking action against content violations, have been widely studied in CSCW and HCI research [11–13, 35, 36]. They often frames moderation as an important part of platform governance that shapes who can participate and what kinds of speech can remain visible [37]. Content moderation affects users primarily through punishments, ranging from content removal to user account suspension [38]. They vary across platforms and are often invisible and unclear [39]. In addition, moderation experiences can be uneven across communities and identities, which makes questions about fairness and bias central to how users experience and respond to moderation [35].

Especially when it comes to politics, the high-stake scenario when maintaining stability, social media are sometimes regulated by certain authorities to enforce censorship on political posts such as domestic or foreign policy. This is usually state-level censorship that prioritizes the nationalization of the internet to restrict access to the internet and monitor sensitive issues [40–42]. Unlike regular content moderation, this censorship is enforced through legal frameworks and technical infrastructure [43, 44], where users have little choice, such as switching to more "free" platforms, as all platforms are subjected to the national infrastructural control. Such censorship is unstable and ever-changing [45, 46], aiming at silencing collective expression (as King et al. [27]) to limit collective action (as Perry [47, 48] and maintain social stability [49]. Empirical works on Chinese platforms also documents censorship as everyday deletion: it is a proactive and automated filtering mechanism, where politically sensitive terms can be associated with higher deletion rates, and accounts can also be closed for usage of politically sensitive terms [50, 51].

Even so, some social media users are not aware of the censorship, and for those who are aware of such censorship, some users are not against censorship and consider it as normal, their overall diverse attitude is heavily dependent on their sociotechnical and ideological background [52–54]. This variation matters because it shapes whether users interpret moderation events as "normal platform governance," political censorship, or personal targeting, which in turn affects their willingness to continue participation [14, 55, 56].

Moreover, the users, especially active users, are actually adjusted to co-evolve with the moderation and censorship. They learn to posit themselves in a less vulnerable stance by navigating the gray areas of information actually permissible by censorship, adopting a collaborative effort that involved both friends and strangers [54, 57]. Such

collaboration exists in user-driven algorithm auditing, where users often take initiatives and work together to surface harms, with different roles such as hypothesizing, data collection, amplification, and contextualization [58, 59]. Our research extends such collaboration into an implicit auditing based on folk theories of the borderline of moderation and censorship, and focus on the reacting mechanism of such collaborative expression for political satire.

In summary, previous literature has examined social media as a space for political expression and ideological clustering [4, 33, 41]. It has also explored how users respond to platform moderation or perceived bias [37, 60]. However, fewer studies have examined collaborative forms of political expression when moderation operates as a systematic regulation of political speech aligned with governmental policies.

In particular, much prior work emphasizes either (a) platform-level governance and moderation outcomes (e.g., removals and sanctions), or (b) individual-level evasion and self-censorship, with less attention to how political meanings are collaboratively built and maintained under systematic political censorship in everyday platform interaction. In this paper, we focus on the full mechanism of collaborative expression under political censorship: why political triggers lead users to shift from direct speech to deniable forms, how such meanings are encoded, how they are relayed and refined through interaction among strangers and weak ties, and what outcomes this produces for participation and collective expression.

2.2 Reaction to Content Moderation

HCI research shows that opaque platform systems shape not only what content is available, but also how users manage visibility, self-presentation, and a sense of being seen on the platform [61–64]. From this perspective, online social presence is not only about speaking, but about staying present in a way that is both visible enough to matter. Therefore, in face of inconsistent or difficult-to-diagnose content moderation [65], to maintain appropriate online social presence, users react differently, depending on the extent of governmental regulation and user sociotechnical status. In HCI, these reactions are often framed as adaptive self-presentation under uncertainty: users continuously adjust what they post, how they post, and where they post as they learn the platform's shifting boundaries [66, 67].

Studies connect users' understandings of moderation to downstream behavior: when people do not understand why content was removed, they are more likely to feel the process is unfair and may disengage; when guidelines or explanations are clearer, they are more likely to keep participating [37]. Beyond explanations, research shows that users actively contest moderation decisions [68–70].

When guidelines are unclear, users might rely on personal "folk theories" [60, 62, 65, 67, 71], based on their known platform engagement practice via first hand experience and social interactions, to understand the opaque and mysterious content moderation algorithm. Their folk theories of the curation algorithm will impact their strategies against the platform moderation [72]. Users reason about how platforms classify or suppress certain identities, and they may intentionally perform for the algorithm or resist such classification [73].

For nontransparent algorithm changes made by the platforms, active users may incite resistant backlash. CSCW work on algorithmic change highlights that backlash is often driven by expectation violation, and that folk theories provide the vocabulary users use to justify resistance such as "fairness", "authenticity" [60]. Because platforms constantly evolve, users may need to repeatedly revise their folk theories and decide whether to adapt, limit their participation, or leave—especially when visibility is tied to self-presentation goals [67].

For more intentional algorithm-driven moderation, HCI/CSCW research documents a range of tactics users adopt. Rather than a static "list," these tactics are often used together and adjusted over time as users receive feedback:

(1) Self-censorship, that is deliberately preventing oneself from posting and expressing to better control presentation of self [74, 75], are widely adopted for avoiding negative consequences especially when audiences are unknown and surveillance feels plausible [11, 66]. CSCW work shows that self-censorship includes both "last-minute" withholding and broader decisions to avoid certain topics or audiences [74–76]. In settings with state blocking and censorship, users adopt a general stance of caution about visibility, applying self-censorship, shaping information cocoon of what kinds of content they are willing to contribute online [77].

(2) Encoding and lexical variation. Lexical variant, including modifying letter compositions or deliberate misspellings, may bypass the content moderation and gain participation and support [78]. Studies show that communities can systematically develop lexical variations in response to moderation pressure, and that these variants can restructure participation and attention within the community [78]. Related work on moderated domains (e.g., misinformation communities) similarly documents how users treat lexical variation as a form of folk theorization: they invent substitute keywords, repurpose features, and iterate tactics as enforcement shifts [79]. In China specifically, work on "sensitive word culture" documents systematic recoding practices (euphemisms, homophones, abbreviations) that evolve alongside keyword filtering and become part of everyday digital cultural production [80].

(3) Satire, or joking. One useful entry point is to treat joking as a practical way to lower political risk while still enabling shared understanding. Research suggests that some satirical tactics are more likely to survive censorship, such as "parodic satire" that mimics official discourse and places it into new contexts, allowing critique without producing an easily identifiable symbol of resistance [25, 81]. At the same time, interviews with Chinese political satirists show that strengthened censorship increases the personal and creative cost of humor and can push creators toward self-censorship and strategic retreat [82].

Taken together, these studies suggest that users' responses to moderation are best understood as a co-evolutionary process: users form folk theories, experiment with tactics, learn from sanctions and peer discussion, and iteratively re-design their performances to remain present. We call this strategic outcome *expressive engineering*: the practical work of designing expressive forms that satisfy multiple constraints at once (e.g., remaining readable to intended audiences, maintaining visibility, and lowering enforcement risk) [65, 67, 73, 83].

In this paper, we use expressive engineering to move beyond a catalog of "tricks." Instead of explaining satires as a set of tactics, we examine how satire operates as an interactive process—how jokes are produced under constraint, how they spread through platform-native interaction (e.g., comments), and how shared understanding is collaboratively achieved while avoiding explicit interpretation that could increase risk. We explore this space by focusing on the expressive engineering in such resilience, elaborating on the strategies.

3 Methods

3.1 Research Design

This study employs a qualitative research design to explore how young Chinese social media users utilize humor and satire to engage in public discourse, and how these expressions are influenced by platform functionality, audience composition, and perceived risk. Given that this study focuses on participants' interpretation processes, audience management strategies, and risk assessments—elements that cannot be directly observed from tracking data—we conducted semi-structured interviews based on examples of social media content provided by participants.

3.2 Participants

We employed a multi-entry snowball sampling strategy to recruit participants. Initial invitations were sent to various social circles, including university networks, professional networks, and interest groups, to reduce sample homogeneity and avoid overrepresentation by a single group. The recruitment announcement described the study as "Humor and Public Discourse Expression on Social Media," deliberately avoiding explicit politicized language to minimize self-selection bias and reduce perceived risk.

Interested participants were required to complete a brief screening questionnaire (3-5 minutes), which included demographic information, frequently used platforms, typical interactive behaviors (browsing, liking, commenting, forwarding, content creation), main interaction scenarios (public information flow and familiar networks), and willingness to share 4-6 examples of expressions during an interview. Participation was entirely voluntary; potential participants were asked to contact the research team directly; the research team would not share any third-party contact information.

We recruited 26 participants (P01–P26), all of whom had actively used at least one mainstream Chinese social media platform (e.g., Weibo, Bilibili, Douyin, Xiaohongshu, Zhihu, or WeChat Moments) within the past three months. To capture the diversity of practices rather than statistical representativeness, we employed maximum variation sampling logic and dynamically adjusted the recruitment strategy to balance key dimensions, including platform usage, interaction intensity (low, medium, high), main expression scenarios (public or familiar networks), educational and professional background, and self-reported interest in public affairs. When a dimension became overly concentrated, subsequent recruitment prioritized the variability of other dimensions.

All participants were anonymized, and we did not collect any identifiable personal information. Detailed characteristics of the participants are provided in Appendix 1.

3.3 Ethics

Given the potential for sensitive information in discussions of public issues within a digital media environment, we took particular care to minimize the risks to participants. This research protocol has been reviewed and approved by [name of the ethics committee requiring anonymity]. All participants signed informed consent forms prior to the interviews and were informed that they could skip any questions or withdraw from the research at any time without penalty. To mitigate potential exposure risks, we did not collect real names, account identifiers, or any directly identifiable personal information. Participants were assigned anonymity codes (P01–P26), and all interview materials were anonymized during transcription. In the report excerpts, we removed or summarized information that could reveal specific individuals, places, or events. Any example expressions shared by participants were considered contextual descriptions, not traceable data points. Interviews were conducted in a private setting (online or offline), and recordings were stored on encrypted devices accessible only to the research team. Data was used solely for research purposes. Given that participants discussed practices formed under perceived risk conditions, we carefully designed questions to avoid inducing explicit political statements and allowed participants to describe their experiences within a comfortable range. Regardless of the depth or specificity of the answer, we offer a reward of 100 RMB (approximately US$14).

3.4 Interview Process

We conducted 26 semi-structured interviews. Each interview lasted 70-90 minutes and was conducted via an encrypted video or audio conferencing platform. Audio recording was done with the explicit consent of the participants. All interviews were transcribed verbatim. During transcription, we summarized potentially identifiable contextual details (e.g., specific cities, institutions, events) to prevent traceability.

The core objective of our semi-structured interviews revolved around the analytical framework of "expression structure—motivation mechanism—collaboration process—risk shaping—feedback loop," systematically understanding how satirical expressions operate in constrained environments. The overall logic followed a progressive path "from surface language to deep mechanisms," and used four fixed key phrases (S1–S4) as analytical anchors to ensure structural comparability and mechanism density of the data.

Firstly, a safety briefing was included at the beginning of the interviews for both ethical and authentic reasons, to reduce psychological burden beforehand. The first part emphasizes "rewriteability, paraphrasability, and non-specificity," shifting the focus of discussion from specific political stances to "expression forms and usage processes," thus concentrating the research on pragmatic structure rather than the events themselves.

The second part, "Platforms and Usage Contexts," establishes the contextual conditions under which expression occurs. Jokes are not abstract linguistic phenomena but

rather arise within the distribution logic of specific platforms, visibility structures, and social network relationships. Therefore, it's necessary to distinguish between public visible spaces and familiar spaces, and to differentiate between browsing, commenting, and original content creation behaviors to understand the relationship between expression forms and audience structure.

The third part, "Joke Collection and Lightweight Interpretation," does not presuppose expression categories but avoids the leading effect of researchers' a priori classifications by having participants recall numerous concrete examples. Through a three-question structure of "literal meaning—contextual meaning—understanding mechanism," different expression techniques such as semantic reversal, referential shift, gesture management, and symbol buffering can be initially revealed. Four key jokes (S1–S4) are then forcibly selected to repeatedly explore the same expressions in subsequent interviews, forming in-depth, vertical rather than scattered, horizontal material.

Part Four, "Why Use It," aligns expression types with motivational structures. It begins by asking open-ended questions about overall motivation, then requires ranking and alignment to specific jokes, to distinguish between abstract reasons and concrete practices. In particular, by probing "What would happen if we didn't use jokes?", it reveals the risk calculations, emotional regulation, and social presence needs behind the expression, thus avoiding the misinterpretation of jokes as mere humor.

Part Five, "How Use It," focuses on the production and collaborative mechanisms of expression. It emphasizes triggering contexts, comment relay, fallback plans, corrections, and handling of misinterpretations, aiming to reveal that jokes are not individual monologues but rather meaning-building processes gradually achieved through group interaction. By having participants recall the process of "everyone adding a line," it captures micro-interactive practices such as collaborative repetition, escalation testing, and risk probing, thereby understanding how "those who understand, understand" is continuously maintained.

Part Six, "Safety and Risk," shifts to the environmental shaping logic of expression. By distinguishing between systemic and interpersonal risks, and by inquiring about risk assessment and escalation order, the study identifies how participants reason about audience structure and how expression changes its form as risk expectations shift. The question of "whether the safety line was rewritten" aims to reveal the dynamic adjustment process of expression strategies.

Part VII, "Consequences," serves as an observation window for the feedback mechanism. By simultaneously asking participants to recall both positive and negative experiences, the study can analyze the dual effects of teasing: on the one hand, it brings resonance and belonging; on the other hand, it may lead to misinterpretation, conflict, or cynicism. More importantly, asking "whether the expression style was changed afterwards" can construct a cyclical structure of "motivation—expression—consequence—readjustment," thus avoiding viewing teasing as a static phenomenon.

Overall, this questionnaire adopts the design principle of "case-driven—structured probing—mechanism induction." It maintains openness, allowing expression types to be generated from the data, while ensuring analytical depth through the S1–S4

structural anchoring. Its goal is to reveal how language is engineered in constrained environments, how it is maintained by groups, and how a continuous strategic balance is achieved between risk and expression.

3.5 Data Analysis

This study employs a grounded theory-based thematic analysis, combined with process tracing and mechanism construction approach, to systematically code 26 semi-structured interview transcripts and approximately 104 samples of joking/satire expressions. The analytical goal is not to semantically classify joking, but rather to reveal its expressive structure, motivational mechanisms, collaborative processes, and risk-shaping logic in constrained environments, thereby answering RQ1–RQ4.

3.5.1 Open-ended Coding

After the verbatim transcripts of the interviews were completed, the two researchers first conducted a line-by-line reading and preliminary annotation. This stage did not presuppose a theoretical framework, but rather focused on conceptually deconstructing the participants' original expressions. We focused on four types of information units: Expressive structural cues (e.g., irony, feigned ignorance, historical allusion, quotation marks, emojis, formations, etc.) Motivational statements (e.g., "afraid of being deleted," "just venting," "wanting to be seen," "it's safer if everyone says it this way," etc.) Interaction process descriptions (e.g., "someone responded in the comments," "testing the waters first," "someone pointed out I was too direct," etc.) Risk and consequence descriptions (e.g., "being misunderstood," "more cautious after the post was deleted," "withdrawing from WeChat Moments," etc.) In this stage, we generated 240 initial codes, encompassing both expressive techniques (e.g., "tone buffering") and emotional and risk assessments (e.g., "anxiety about single-point exposure"). The coding granularity was kept as fine as possible to avoid premature concept merging.

3.5.2 Axial Coding

Building on the open coding, the research team aggregated and restructured concepts using a repeated comparison method. We reconnected scattered concepts along four main lines: "expression—motivation—risk—interaction," forming more explanatory mid-level categories.

Encodings such as "quotation marks," "dog emoji," "mute," and "..." are grouped into tone buffering mechanisms. "I'm not sure how to comment," "it is said," "paraphrasing," and "this is just my personal opinion" are grouped into denial posture structures. "Everyone +1," "Let's form a line," and "queue up" are grouped into collaborative repetition mechanisms. "Leave a memento before deleting," "Observe the wind direction," and "start lightly and then heavily" are grouped into risk probing and escalation paths.

Simultaneously, we cross-referenced motivational codes with expression structures to identify their structural correspondences. For example, the motivation of "risk avoidance" is highly co-occurring with "denial posture," "tone buffering," and "referential transfer"; the motivation of "group security" is highly correlated with

"collaborative repetition." Through this cross-matrix, we gradually formed a mapping relationship between seven types of satire structures and six core motivations.

3.5.3 Selective Coding

Based on axial coding, we constructed an explanatory framework around the research question, ultimately forming a three-tiered structure: Expression Structure Layer: Seven types of teasing (semantic reversal, denial posture, referential shift, tone buffering, cynical response, collaborative repetition, and anecdotal expression); Motivation Layer: Risk aversion, emotion regulation, social presence, cognitive distance, group safety, and long-term psychological defense; Process Layer: Trigger-Encoding-Relay-Stable Collaborative Cycle Mechanism

Through selective coding, we integrated the aforementioned categories into a dynamic model: Motivation-Driven Expression Coding → Collaborative Relay in the Comment Section → Positive and Negative Consequences → Rewriting the Risk Threshold → Influencing the Next Round of Expression. This cyclical structure becomes the theoretical thread of the subsequent analysis.

3.5.4 Process Coding

We specifically coded the narratives involving the interaction process. Specifically, this includes: Trigger types (hot topics, narrative discrepancies, template discourse); Exploration paths (starting lightly and gradually escalating); Correction mechanisms (reminding of overly direct statements, suggesting rewriting); Retreat methods (deleting posts, migrating, remaining silent). By reconstructing these codes chronologically, we identified the dynamic path of "expression engineering" and how strategies progressively escalate as risks increase.

3.5.5 Negative Cases and Boundary Testing

To avoid over-theorization, we actively sought cases that did not conform to the main pattern. For example: A few participants still chose cynical expressions in low-risk scenarios; Some jokes were misinterpreted within the circle; Some people explicitly refused the collaborative rereading mechanism. Through the analysis of these negative cases, we corrected some overly linear explanations, adjusting the model to a probabilistic rather than a deterministic mechanism.

3.5.6 Coding Reliability and Consistency

To improve the rigor of the analysis, the two researchers independently coded 30% of the interview materials and reached a consensus through discussion. We did not use the statistical consistency coefficient as the sole criterion, but instead calibrated it through "conceptual boundary clarity" and "classification interpretability." The coding conflicts mainly centered on the distinction between "semantic reversal" and "joke-like expression," ultimately achieving unification by introducing the criterion of "whether semantic tension is the core mechanism."

Given the sensitive context involved in the research, the research team continuously reflected on their stance during the analysis process, avoiding overly romanticizing

satire as "resistance" or simply categorizing it as "a joke." We emphasize its duality: it is both an expressive engineering effort in a risky environment and a long-term psychological defense mechanism.

4 Results

4.1 Vernacular Steganography: A Typology of Satirical Expression

Based on the analysis and coding of approximately 104 jokes, we identified seven recurring expression structures in constrained environments. Drawing on boyd's concept of "social steganography" [84, 85]—where messages are encoded to be accessible to intended audiences while remaining opaque to surveillance—we categorize these expressions not merely as rhetorical devices, but as a set of *steganographic affordances*. Unlike traditional steganography which hides the existence of the message, these expressions hide the *intent* of the message. Different categories differ in expression intensity, level of denial, and collaborative potential, but collectively constitute a vernacular encryption system that can operate in risky environments.

4.1.1 Semantic Reversal and Exaggerated Irony

This type of expression achieves semantic reversal through overly positive, extremely exaggerated, or significantly elevated tones. Its core mechanism is not direct negation, but rather the creation of semantic tension, causing a cognitive conflict between the literal meaning and the contextual reality, thus automatically deriving the opposite meaning. The advantage of this type of expression lies in its high degree of plausible deniability: completely positive at the textual level, and critical at the contextual level.

For example, one participant used the phrase:

"They really won again." (P13)

This expression borrows the narrative framework of "victory," but in a specific context it constitutes irony, implying the gap between the actual outcome and the official positive narrative.

Another participant similarly wrote:

"Let me just say, this move was truly high-level." (P7)

Here exaggerated praise communicates denigration. Readers do not understand the criticism through negative words, but rather identify the indirect criticism through excessive praise.

This category is particularly common in constrained environments because it allows the speaker to convey their stance without explicitly using critical language, avoiding direct confrontation while quickly establishing implicit understanding within a consensus-building group.

4.1.2 Denial Stance and the Construction of Retreat Space

Unlike the first type, the core of the second type of expression lies not in semantic reversal, but in stance management. Its goal is not "what to say," but "how to retain space for what is not clearly stated." This type of expression, through feigning ignorance, ambiguity, disclaimers, and responsibility transfer structures, pre-sets a retreat path, allowing the expression to be redefined when questioned.

For example, participants used expressions such as:

"I don't understand, but I'm deeply shocked." (P26)
"My parrot is lost, and I disagree with my parrot's viewpoint." (P7)
"I'm not sure how to evaluate." (P8)

These expressions superficially present confusion or uncertainty but in context often communicate absurdity or criticism. They weaken the subject's responsibility at the syntactic level, keeping the stance revocable.

This category constitutes the foundational layer of deniable expressions. If the first category addresses "how to criticize without explicitly stating it," the second addresses "how to back down when questioned."

4.1.3 Referential Shift and Metaphorical Substitution

The third type of expression achieves referential shift through time, space, or historical frameworks, embedding contemporary criticism into an alternative narrative structure. Its mechanism lies in: by alluding to the past to satirize the present or blurring spatial boundaries, the expression is detached from the direct object, thereby reducing literal risk.

For example, participants used phrases such as:

"quite reminiscent of Chiang Kai-shek's bandits." (P12)
"four generations of quack doctors, good for nothing." (P15)
"The Qing Dynasty hasn't fallen yet? I'm truly impressed." (P14)

These expressions use historical, political, or family metaphors to allude to contemporary situations. They do not name the actual subjects directly, but convey meaning through shared historical context.

The function of this type of expression is to establish cognitive distance: the criticism does not occur "here and now," but is reflected in "history," "the other," or "fictional text." The expression thus becomes indirect, but does not lose its direction.

4.1.4 Tone Buffering and Symbolic Insurance Mechanism

The fourth type of expression does not bear the primary semantic burden, but rather serves as a pragmatic "insurance layer" attached to the text. Through punctuation, emoticons, abbreviations, or linguistic breaks, the expression is repositioned as a joke, a vague comment, or a mild gesture.

Examples include:

"..." (P1)
"dog head." (P25)
"666. This is really 6." (P17)

Here *"666"* or *"6"* stands for "impressive" in Chinese and may express admiration or sarcasm depending on context.

These symbols are often used in conjunction with expressions of the first or second category, forming a multi-layered deniability structure.

The key function of the fourth category is to weaken certainty, making the expression difficult to categorize unilaterally in an algorithmic or human censorship environment.

4.1.5 Cynic Responses and Emotionally Dull Expressions

The fifth category of expressions responds to public events through indifference or habitual gestures, rather than creating semantic tension. Its goal is not to reinforce criticism, but to reduce emotional intensity.

For example, participants used expressions such as:

"I'm used to it." (P22)
"It's nothing new." (P24)
"What else can I do?" (P16)

These expressions weaken the impact of events through de-emphasis. They do not emphasize stance, but rather "anticipation" and "normalization," transforming the abnormal into the everyday.

In risky environments, cynical responses serve as both a psychological defense and a public de-escalation strategy. However, long-term use may alter the discussion ecosystem, shifting public issues from argumentative to symbolic.

4.1.6 Collaborative Repetition and Group Confirmation Mechanisms

The sixth type of expression does not add new semantics but achieves a scale effect through repetition, chain reactions, or standardized formats. Its mechanism lies in embedding individual expressions into group postures, transforming risk from a single-point exposure to a distributed existence.

For example, participants used expressions such as:

"+1" (P3)
"Agree with the comment above." (P14)
"Stay in line." (P25)

Here "stay in line" means to follow the original poster by using the same words.

Repetition not only strengthens consensus but also reduces individual salience.

This type of expression often combines with irony (Section 4.1.1) or metaphor (Section 4.1.3), causing templated expressions to spread rapidly within the group.

4.1.7 Emotional Translation through Jokes and Disaster Humor

The seventh type of expression transforms serious issues into plots, scenes, or program effects, reconstructing reality through an entertaining framework.

For example, one participant used the phrase:

> *"choosing the method described longest in history books."* (P10)

This expression dramatizes complex decision-making into a textbook plot, shifting the speaker from participant to narrative observer. This type of expression reduces the intensity of conflict and provides cognitive distance.

Unlike the first type of expressions (Section 4.1.1), the seventh type emphasizes narrative and textualization, allowing reality to be "watched" rather than directly confronted.

4.2 RQ1: Why use satire/deconstruction to discuss politics – Negotiating the Tension Between "Speaking Out" and "Surviving"

Through a discussion of specific meme examples, it is shown that our users do not simply "love using memes," but rather engage in continuous strategic calculations between emotional intensity, social relationships, and risk expectations. From the perspective of Communication Privacy Management (CPM) theory [86], participants are constantly navigating the dialectical tension between the need for revelation (expression) and the need for concealment (safety). The categories of memes serve as boundary regulation mechanisms that allow users to maintain this precarious balance. An observable structural correspondence exists between motivations and categories. Specific details are shown below.

4.2.1 Risk Aversion: Satirical Expression is an "Expressive Engineering"

In summary, risk aversion is not a byproduct of satire, but rather its structural premise. In most online discourses, the sarcastic structure of political commentary is directly caused by restrictions on social media platforms and users' proactive concerns about their own safety.

For example, one participant explained:

> *"If I could just say, 'Our current system is absolutely rotten, our economy is absolutely rotten, my job is absolutely rotten,' why would I say the opposite—'The economy is stable and improving'? Isn't that something they (social media platforms) don't allow? Saying that gets my account banned and warned."* (P1)

This motivation primarily activates three types of sarcastic structures: the second type, "denial stance"; the fourth type, "tone buffer"; and the third type, "reference transfer," constructing deniability at different levels.

Denial-based expressions reduce their intensity through syntactic weakening and shifting of responsibility, keeping the stance revocable. For example:

"When I saw that video mentioning that China's GDP last year actually surpassed that of the United States, you know, the blogger meant that, according to his calculation method, well, all I can say is, I don't understand, but I was deeply shocked." (P26)

Tone-buffering expressions add a "safety shell" to the text surface through punctuation, censorship, and semantic distortion, weakening the certainty of algorithmic or human interpretation. For example:

"Sometimes algorithms are quite stupid. For example, if you type 'economic downturn,' you might get a warning, but if you type 'economic' followed by a downward arrow, it might not get a warning, but everyone can understand that arrow." (P18)

Referential shifting expressions avoid direct naming through spatial or object substitution, thereby reducing aggression and targeting. For example:

"Our financial situation here is so bad that the bus system has to borrow money from temples to pay its employees' salaries. The news keeps saying the municipal finances are good. What can we say? We can only say that our Celestial Empire is invincible. You don't understand, this is the strength of the great Eastern power." (P23)

These three types of structures achieve risk diversification at the grammatical, pragmatic, and referential levels, respectively, collectively forming a multi-layered expressive protection mechanism.

Functionally, it is a kind of "risk engineering." When constructing sentences, the speaker presupposes possible paths of questioning and scrutiny, and reserves a way out by embedding denial clues. Risk avoidance thus becomes a crucial driving force for meme selection, enabling the expression to convey a stance signal while also being reinterpreted as a joke, paraphrase, or ambiguous opinion when necessary. In constrained environments, this multi-layered denial structure constitutes a key condition for the continued existence of banter.

4.2.2 Emotion Regulation: Transforming "Anger/Powerlessness" into a Tolerable Linguistic Form

In summary, the core of this motivation is emotion governance. In public discourse characterized by high uncertainty and low sense of control, directly expressing anger often carries a high risk of conflict and significant psychological strain.

For example, one participant noted:

"You know, expressing dissatisfaction on social media is risky, very risky, and not just because of the risk of power authorities. I feel uncomfortable when my anger is seen by people I don't like; I'm afraid my anger will become their laughingstock." (P2)

Complete silence, on the other hand, signifies emotional repression and social withdrawal. Another participant explained:

"Given the situation I'm facing, this messed-up reality, if I don't vent, I feel like I'll suffocate sooner or later." (P7)

Therefore, irony becomes a middle ground: it allows for the externalization of emotion while avoiding direct conflict.

This motivation primarily occurs in three types of irony: semantic reversal, anecdotalization, and cynical responses. They all serve an emotional regulation function, but in different ways.

First, semantic reversal creates absurd tension, transforming anger into irony. For example, participants described expressions such as:

"We won again!"
"We're so tired of winning!"

P2, P4, P7, and P11 all emphasize that this is a process of emotional translation. One participant explained:

> *"These expressions are literal narratives of victory, but their overly positive tone is completely at odds with real-life experience; anyone would find it unbearable."* (P11)

Secondly, by reconstructing the narrative framework through anecdotes, the speaker shifts from a participant in the event to an observer, gaining psychological distance. For example:

> *"When I was arguing with a blogger who was mindlessly patriotic, I felt that the other person was completely unreasonable, constantly saying 'Patriotism is never wrong,' seemingly unable to understand what I was saying. I was really angry, so I said, 'You come from Syracuse.'"* (P12)

Finally, cynical responses reduce emotional intensity and limit the drain from continuous engagement. Examples include:

"I'm used to it." (P22)
"What else can I do?" (P24)

One participant further commented:

> *"What else can I do? If I keep arguing with these idiots, I'll drive myself crazy sooner or later. Forget it, I'm not wise enough to argue with them."* (P25)

These three paths, while mechanistically different, collectively accomplish one task: compressing high-intensity emotions into low-risk, low-cost, and shareable units of expression. Therefore, banter is not merely a rhetorical device, but an emotional regulation mechanism that strikes a balance between expression and survival.

4.2.3 Social Statements: Achieving "I'm Here" Without Explicitly Taking Sides

In summary, the core motivation for "being here" is not about expressing a specific stance, but about managing one's social position. Participants seek a balance between "being seen" and "not being overexposed." For example:

> *"I still hope to be seen, you know, a kind of feeling similar to leaving a trace of my existence, but this trace can't be too obvious, otherwise it's unsafe."* (P13)

Through two-way readable ironic expressions, metaphorical structures of referential shift, and collaborative repetition of group embeddings, individuals can maintain visibility in public discussions while avoiding being fixed as a single-sided subject.

Examples include:

"Open up the scope!" (P11)

Another participant noted:

"Whether it's saying the opposite of what you mean or the opposite of what you mean, I'm here, I exist here, and they can't do anything about it." (P17)

Referential shifts such as:

"the familiar recipe." (P13)
"history always repeats itself." (P15)

Similarly create interpretive flexibility. One participant explained:

"Anyway, what I'm saying is that the Qing Dynasty's system was really bad, the Qing Dynasty's leaders were really bad. Who was the Qing Dynasty? Who were the Qing Dynasty's leaders? I've said it, I exist here, and future generations might see me in history books." (P16)

Collaborative repetition further embeds individual expression into a collective structure. For example:

"+1."
"Let's form a line!"

Another participant described the motivation behind such participation:

"Since we're here, in the spirit of joining in the fun, we can't just leave without saying anything." (P1)

These structures collectively constitute a low-intensity, adjustable participation mechanism: individuals maintain public visibility through identifiable yet deniable pragmatic signals rather than through explicit positional declarations.

Functionally, this transforms "status declaration" into "signal transmission," making participation a social behavior with controllable intensity. In risky environments, this light-touch visibility maintains social connections while avoiding identity solidification and positional lock-in.

4.2.4 Cognitive Distance and Humorous Self-Salvation: Transforming Grand Issues into Processable Narratives

The core motivation of "cognitive distance" is not about escaping reality, but about regaining narrative control. Faced with large-scale, structurally complex public issues that are difficult for individuals to intervene in, participants construct psychological and semantic buffers through "reference transfer" and "joke-like expression."

Firstly, regarding reference transfer, some participants utilize historical metaphorical expressions such as:

"The Qing Dynasty hasn't fallen yet?" (P12)
"You're the successor to the Qing Dynasty..." (P14)

By shifting the timeline, the direct sense of oppression from reality is reduced. One participant further explained:

> *"I only said that the post-Qing dynasty was like this: the system was flawed, the economy was flawed, everything was flawed. As for who the post-Qing dynasty was, whoever felt guilty was the one."* (P14)

In this case, P14 changed the subject of the complaint from social reality to "the post-Qing dynasty," which has a clear derogatory connotation. Other participants used literary references, for example:

> *"1984 is a novel, not an instruction manual."* (P21)

Such references embed contemporary reality into an existing narrative framework through cross-textual comparison, allowing meaning to emerge through analogy. One participant elaborated:

> *"However, the author wrote 1984 to warn future generations. Unfortunately, rulers throughout history have taken the plot of intrigue and trickery in the novel, which was originally intended as a warning, as an instruction manual."* (P21)

Secondly, regarding anecdotal structures, participants further translate complex decisions into "plot choices" that can be viewed, discussed, and evaluated. For example:

> *"choosing the method that will be described longest in history books."* (P10)

The participant further explained:

> *"Well, regarding these political decisions, I can only say that history books can write very, very long analyses, and whether they are positive or negative depends on how future generations interpret them."* (P10)

In this case, P10 expressed dissatisfaction with certain policies, arguing that these policies could have significant long-term impacts and might eventually occupy substantial space in historical narratives.

In these expressions, real-world problems are transformed into texts, history, or stories, and the speaker shifts from a passive recipient of reality to an interpreter within the narrative. Mechanistically, this type of expression does not directly intensify criticism but rather transforms "uncontrollable reality" into "discussable text" through semantic abstraction and narrative reconstruction.

Cognitive distance thus becomes a self-protective strategy: it avoids both direct confrontation and complete silence, allowing individuals to maintain a minimum level of interpretive power and subjectivity in the face of grand issues. This motivation emphasizes a sense of cognitive control. Through metaphor, analogy, and textualization, individuals regain interpretive power in the face of grand issues.

4.2.5 Group Safety: Risk Dispersion Through Collaboration

From a motivational perspective, the group safety motive stems from an individual's high sensitivity to "single-point exposure." In high-risk or highly uncertain discussion environments, expressing a clear stance alone means taking on identifiable risks, whether from platform censorship, reporting mechanisms, or misinterpretations by

audiences across different demographics. Participants frequently mentioned (23 out of 26) that on sensitive topics, "it is too conspicuous for one person to speak." When expression becomes repetition by multiple users, individual visibility is reduced. As one participant explained:

> *"Everyone says so."* (P4)

Structurally, this also disperses responsibility, preventing expression from being concentrated on a single speaker. The motivation here is not to strengthen the stance itself, but to avoid becoming the focal point, thereby reducing the risk of personal exposure while maintaining participation.

In terms of expression, this motivation often appears in the sixth meme structure we identified, "collaborative repetition." Examples include:

> *"The template is here."* (P11)
> *"Queue up. Queue up."* (P15)
> *"Uniform format."* (P17)
> *"This really is the standard answer."* (P19)

Short phrases such as *"+1"* (P1, P3, P5, P22) or *"Agree with the comment above"* (P14) function similarly. These expressions usually add little new semantic content. Instead, repetition creates a scale effect through chain-like structures or formations.

At the pragmatic level, repetition embeds individual expression within a collective stance. Risk shifts from a single comment to a distributed expression.

Furthermore, this repetition structure often combines with the first type of "semantic reversal." For example, the ironic phrase *"They really won again"* (P14) may appear repeatedly in a queue-like structure, breaking a critical statement into multiple low-intensity units. Through this collaborative repetition, risk is redistributed within the group, maximizing individual safety.

Group safety motivations therefore emphasize risk dilution. They do not introduce new information but instead produce scaled expression. This scale carries both psychological and structural significance. Psychologically, statements like *"Everyone says so"* reduce individual anxiety. Structurally, the message is distributed across many similar statements, reducing the salience of any single comment.

4.2.6 Long-Term Psychological Defense: Cynicism as a Survival Strategy

From a motivational perspective, long-term psychological defense emerges from the chronic depletion associated with continuously risky environments. When individuals repeatedly experience post deletions, arguments, misunderstandings, or identity exposure in public discussions, expression becomes a high-cost activity rather than a simple emotional outlet. As one participant explained:

> *"In a continuously high-pressure context, maintaining anger or sustained serious participation is unsustainable."* (P6)

Unlike immediate emotional regulation, this shift is not a reaction to a single event. Instead, it reflects a postural adjustment from wanting to express one's stance to

lowering expectations about the impact of participation. Another participant described this feeling as:

> *"Many times, I really feel a strong sense of powerlessness, as if I've lost all expectations, so I just let things be."* (P19)

In this sense, cynical expression becomes a survival strategy. By lowering emotional investment and reducing the intensity of one's stance, individuals protect themselves from long-term psychological depletion. This defensive posture does not necessarily indicate indifference, but rather adaptation to a persistently uncertain environment.

In terms of expression, this motivation appears in what we identify as the fifth meme structure, *cynical response*. Examples include:

"It's all the same." (P6)
"Don't take it too seriously." (P9)
"There's nothing new under the sun." (P9)
"Nothing's surprising." (P19)
"Just do your best." (P19)

These expressions rarely contain explicit criticism or strong irony. Instead, they weaken the perceived importance of events through a flat tone and familiar framing. Unlike semantic reversal or joke-like expressions, cynical responses avoid humor or dramatic tension. Rather, they frame events as predictable or unsurprising through emotional distancing and de-emphasis.

At the pragmatic level, such expressions transform public events into everyday noise, allowing individuals to limit cognitive and emotional investment. Over time, this pattern may reduce the intensity of discussion, gradually shifting public interaction from argumentative exchanges to a more detached communicative climate.

4.3 RQ2: How These Jokes Are Produced, Disseminated, and Collaboratively – Establishing "Common Ground" through Ambiguity

Regarding each example joke, participants often do not first discuss "what the sentence means," but rather "where it appeared," "in what atmosphere," and "why everyone is suddenly using it at this time." This suggests that in the participants' view, a joke is not a product of a single individual's expression, but rather a collaborative process triggered by an event, processed by the group, and gradually "anchored" through interaction. We analyze this process through the lens of Clark [87]'s common ground theory. Usually, common ground is established through explicit confirmation [88]. However, in this context, participants engage in a form of "collusive grounding," where mutual understanding is confirmed not by clarity, but by the correct usage of ambiguous codes. For example, one participant explained:

> *"A joke out of context is meaningless. It's somewhat similar to, in mathematics, ignoring monotony in isolation; you know, you have to put that sentence in the right context to achieve the corresponding effect."*

Based on our continuous questioning of "location of occurrence—meaning inference—others' responses—rewriting and superposition—understanding confirmation," we synthesize these user experiences into a four-stage framework of collaborative sense-making: triggering—encoding—relay—consolidation. These four stages together explain why participants believe the same sentence is "instantly understood" by insiders but "unnoticed" by outsiders, and why "interpretation" itself is considered a high-risk behavior.

4.3.1 Perceived Triggers: Events as "Tipping Points" for Expression

Participants described that most of their jokes did not stem from a pre-planned discussion of a particular topic, but rather from an immediate reaction to a sudden situation. For example, one participant explained:

> *"Let's say, when I hear Zhang Weiwei again praising China and criticizing the US, I'll say 'China has won again.' But this description criticizes illogical, position-driven, and untrue statements. For instance, when I read history and see a king of an autocratic country telling his people how happy their lives are, I'll also use the expression 'won again.' These are completely different scenarios, but the core trigger is exactly the same."*

The first type of trigger identified by participants stems from the suddenness and high uncertainty of public events. When information is fragmented, accounts differ, and the outcome is unpredictable, individuals find it difficult to express their stance with serious arguments. Instead, they are more likely to use short, reproducible memes to "claim a spot." They describe this state as "Information is too chaotic, so I just put my attitude in one sentence." One participant stated:

> *"When something first happens, it's impossible to explain clearly; writing a long paragraph is useless. Just start with something like 'We won again,' and everyone knows what you mean."* (P13)

Secondly, the second type of trigger comes from the perceived gap between official narratives/media statements and individual experience. Participants often mentioned that certain highly templated, overly positive statements activate the first type of "semantic reversal" memes (such as exaggerated praise, rhetorical victory), because this gap naturally creates irony. The meme here is not about "following the trend," but a way to quickly socialize the sense of disparity. One participant explained:

> *"It's the kind of thing where you see it and think, 'Oh, here we go again,' and then the comments section automatically shows things like 'opens up new perspectives' and 'suggests applying for UNESCO World Heritage status.' People are actually complaining about the template."* (P17)

Then, the third type of trigger comes from the sudden appearance of "highly compressed expression templates" in the comments section. Many participants mentioned that memes are often not created by a particular "opinion leader," but rather accidentally generated by ordinary users testing boundaries in the comments section; once a sentence has the characteristics of being "short enough, easy enough to modify, and

having a fallback option," it will be quickly copied, rewritten, and layered, facilitating its rapid spread. One participant noted:

> *"Many memes aren't invented by anyone; it's just that someone in the comments section says something so insightful that everyone sees it as useful and starts copying and pasting, adding emojis, and rephrasing it."* (P4)

Therefore, rather than saying the triggering process revolves around the "event content," it is more accurate to say it revolves around the participants' "expressive needs": the event provides the fuel, the contrast provides the tension, and the comments section provides the infrastructure for template dissemination. What the trigger generates is not "opinions," but rather expressive templates and interactive entry points that can be leveraged by the group.

4.3.2 Encoding Strategies: Constructing a " Package of Linguistic Cues" for Ambiguity

In the encoding process, participants emphasized that they judged whether an expression was "readable by its peers" not based on straightforward content, but on a set of overlapping clues that together constituted a "Package of Linguistic Cues." For example, one participant stated:

> *"Content used for political satire must be able to fully express what is meant to be said, but not actually said, while also being as precise as possible to avoid inappropriate expression, so that the intended audience can understand it."* (P1)

This set of clues often uses the fourth type of "tone buffer" and the second type of "denial stance" as its basic shell, and then embeds the true meaning using the first type of "semantic reversal" and the third type of "reference shift."

For example, P1, P7, P12, and P16 mention that "quotation marks" and "emoticons" are considered the most commonly used rapid encoding tools. When words are enclosed in quotation marks (such as "advanced"), a semantic gap is artificially created, and the reader will automatically be alerted that "this is not the literal meaning"; when a sentence ends with a "dog head" or "...", the expression changes from a statement to a gesture, reducing aggression and increasing deniability. One participant explained:

> *"Actually, you can tell from the quotation marks/emoticon that it's ironic. If you really take it literally, then you're not the target audience."* (P1)

The second type of "denial stance" is often used to change the position from "I said" to "I didn't explain clearly/I was just relaying." Phrases like "I can't comment," "We dare not say," and "This is only my personal opinion" on P8 serve both to indicate "you know what I mean" within the circle and to provide an escape route to outsiders. Participants call this kind of phrasing "paving the way out first." One participant noted:

> *"It's not that I don't want to say, it's just that I'm putting on a 'disclaimer' first, otherwise you'll be labeled with a single word."* (P18)

The third type of "referential transfer" and the first type of "semantic reversal" often serve as a stronger "semantic encryption layer." For example, participants used phrases such as "a certain major power," "across the sea," "The Qing Dynasty hasn't fallen yet?," "History has been revised," "we've won too much," and "thank you" to embed critique through indirect references or ironic positivity. One participant explained:

> *"These expressions are effective because they require readers to share a minimum amount of contextual knowledge: knowing what has recently happened, knowing certain templated narratives, and knowing the common references to certain historical tropes."* (P14)

Another participant added:

> *"You don't need to say who it is; 'across the sea' is enough for everyone to know you're avoiding that name; 'later Qing dynasty' is obviously an allusion, but if you want to be pedantic, I haven't named names."* (P15)

Therefore, encoding is not as simple as "hiding the meaning," but rather a delicate balance between "readable within the circle" and "debunkable outside the circle": the richer the clue package, the more stable the decoding within the circle; the more ambiguous the clue package, the more difficult it is to trace responsibility externally. The result of the encoding process is a semi-public text that can be collaboratively transmitted: it is neither a code nor direct statement, but a hinting system that allows the group to complete it together.

4.3.3 Relay: Relying on "Collaborative Repetition" to Anchor Meaning

Participants generally believe (18 out of 26) that many jokes are "half-finished products" when posted: they provide direction but do not say everything; the true political meaning is often gradually "solidified" through interaction. For example, participants explained:

> *"I said, without a comment section, there would be no such thing as 'joking.' The essence of 'joking' is community interaction; the content of these jokes is the result of group interaction and creation."*

> *"You know, what one person writes might be profound, but it can't be funny."*

In this stage, the sixth type of "collaborative repetition" is not just a means of dissemination, but also a mechanism for meaning production. Repetition (such as "+1," "Agree with the comment above," "Let's form a group") first creates a sense of "we're together," and also creates the illusion that "more people = safer," making more people dare to join. When the scale of repetition reaches a certain point, the meaning of the joke becomes solidified by group consensus, becoming increasingly difficult to explain as "accidental misinterpretation." One participant noted:

> *"For example, the joke about gratitude is because everyone is using it; the law doesn't punish the masses, and the positive meaning of this meme is very strong, so the risk is very low."*

During the relay process, we observed participants repeatedly mentioning four typical operations: escalation, transfer, correction, and superposition. Escalation refers to adding a sharper implication to the original statement, used to test boundaries. For example, someone added "Suggest applying for UNESCO World Heritage status" under "They really won again," further increasing the irony; or someone added "History always repeats itself" under "The familiar formula," making the reference clearer but still debatable. One participant described this process as:

> *"First a mild comment, then someone adds a harsher one; everyone is testing how far things can go today."*

Shifting the focus is used to mitigate damage. When the risk of discussion increases or the conflict escalates, the comments section will steer the topic towards more relatable metaphors or a more abstract historical framework, making the expression ambiguous again. This shift often uses third-type referential shifts and fifth-type cynical stances (such as "It's all the same," "Nothing surprising") to cool things down. One participant explained:

> *"When the wind seems to be blowing in the wrong direction, people will start saying 'Don't take it too seriously,' 'Just take it with a grain of salt,' which is basically a reminder to tone it down."* (P26)

Correction is another key mechanism. When participants frequently mention it, insiders will proactively remind them "Don't be too direct," "Change the wording," "Add a dog emoji," or "Delete a word." Correction is both a form of group self-censorship and mutual assistance, and it keeps expression within a debatable range. One participant noted:

> *"Someone will comment 'You're being too direct,' which is actually a way of helping you, not insulting you."* (P1)

Overlay reflects collaborative writing: the same template is constantly tweaked by different people—changing synonyms, adding quotation marks, parentheses, muting, replacing with geographical metaphors, or shortening sentences for copy-pasting. Through these tweaks, the meme maintains its recognizability while continuously adapting to new risky environments. One participant explained:

> *"Memes aren't copied verbatim; people will 'modify' them, making them seem more like jokes or casual remarks, but those who understand will still understand."* (P13)

Therefore, the core of the relay stage is not the speed of dissemination, but the collaborative completion of meaning: individuals throw out a direction with half a sentence, and the group uses repetition, upgrading, correction, and aggregation to turn that direction into a shared "readable consensus." As one participant said:

> *"If one person says something that sounds too much like a declaration, it's easy to cause trouble. Everyone adding a sentence in the comments section is like breaking a knife into many pieces; it's less conspicuous."* (P12)

4.3.4 Stability: Maintaining the "Cost of Misinterpretation" through Tacit Consensus

Finally, whether a joke can circulate long-term does not depend on whether it is "funny," but on whether a stable set of interactive norms has been formed within the community to maintain its readability and deniability. Participants repeatedly emphasized that "only those who understand will understand" is not a naturally occurring difference in ability, but rather a boundary work that is continuously produced and maintained. For example, one participant explained:

> *"Those within the circle maintain the semi-public nature of memes through default interpretations, moderate hinting strategies, and vigilance against 'over-interpretation.'"*

Specifically, when external misinterpretations occur (e.g., taking irony literally, or metaphor literally), those within the circle often will not publicly "explain clearly," because explanation itself would expose the implicit meaning, rendering the originally negligible structure invalid. Instead, they prefer to hint in more subtle ways: for example, adding "Those who know, know," or using "..." and a "dog head" emoji to remind "this is sarcasm," or using collaborative repetition to suppress the original interpretation as the mainstream. One participant described it this way:

> *"If you actually explain 'what I'm actually satirizing,' isn't that like confessing? Usually, you just say 'Those who know, know,' letting those who understand align themselves."*

Stability also relies on a collective consensus on the "acceptable intensity of expression": which words cannot be used, which metaphors are safer, which quotation marks/mutes are more effective, and when to withdraw or delete a post. Over time, these practices solidify into tacit understandings, becoming "common sense of expression" within the circle. When expression deviates from these common-sense norms (too direct, too sharp, too easily grasped), correction occurs quickly, keeping the meme sustainable within a deniable range. One participant noted:

> *"Everyone actually has a tacit 'safety line.' Whoever crosses it will be reminded, and if they don't correct themselves after being reminded, someone will say, 'Forget it, don't say it anymore.'"* (P11)

Therefore, the essence of stabilizing the process is managing the "cost of misinterpretation" shared within the community. On the one hand, they need to make the meme clear enough for others to decode; on the other hand, they need to make the meme ambiguous enough for external denial. "Only those who understand will understand" is therefore an organized result of interaction, not individual talent or chance.

4.4 RQ3: How Platform Mechanisms and Risks Shape This Expression – Navigating Risk and Audience

Interviews show that participants' risk calculations rarely stop at single judgments like "whether there is censorship" or "how strict the platform is." Instead, they seem to be making audience structure inferences. For example, one participant asked:

"If I post this, who will ultimately see it? Friends in the same circle, strangers, random people recommended by the algorithm, or will it be screenshotted and forwarded to offline relationships by acquaintances?" (P2)

In the participants' considerations, risk is therefore not an abstract "environmental variable," but a dynamic structure jointly determined by the perceived platform distribution mechanism, visibility settings, and social relationship networks. This reflects the classic challenge of "context collapse" [89], where distinct social audiences (friends, family, censors) are flattened into a single context. However, our findings show that users do not passively accept this collapse; instead, they actively employ the steganographic affordances described above to re-segregate these audiences. It is within this inference framework that expression is shaped into a "low-salience shape" that is deniable, transferable, and collaboratively diffused, closely corresponding to the types of banter summarized in the previous two sections: when risk expectations rise, the second type of denial, the fourth type of tone buffering, and the third type of referential shift appear more frequently; when group cover is needed, the sixth type of collaborative repetition is activated; when the discussion space is more open, the first type of semantic reversal and the seventh type of pun-like expression are more likely to occur.

4.4.1 Conceptualizing Risk: The Tension Between Systemic and Interpersonal Concerns

Participants distinguished between two types of risks of different natures: one is platform risk at the systemic level, and the other is interpersonal risk at the social level. Systemic risks are primarily linked to public information flow scenarios, encompassing consequences such as post deletion, traffic restrictions, alerts, and account limitations. When discussing systemic risks, participants frequently mentioned a sense of "uncertainty and being observed": they may not know where the trigger thresholds are, and it is difficult to predict which words will be identified as sensitive.

One participant explained:

"The most annoying thing is that you don't know which sentence will cross the line. Sometimes the same meaning might be acceptable today but not tomorrow. You'll subconsciously write things in a way that's 'more inappropriate.'" (P3)

Because of the opacity of systemic risks, participants proactively embed withdrawal mechanisms through the fourth type of "tone buffering" (such as quotation marks, emoticons, censorship, and white space) and the second type of "denial stance" (such as "I can't comment" or "This is just my personal opinion"). Quotation marks and censorship are not merely rhetoric, but a deliberate weakening of "system searchability"; disclaimers are not just polite formalities, but a deliberate dilution of "accountability."

Another participant noted:

"It's not that I want to be sarcastic, but if you don't put it in quotes and a dog emoji, you're just asking to be taken seriously." (P5)

In contrast to systemic risk, interpersonal risk occurs more frequently within close-knit networks: social circles, colleague groups, family groups, or local communities. Participants pointed out that interpersonal risks often manifest as misunderstandings,

broken relationships, exposure of real-life identities, and the spread of screenshots, and their harm tends to have a longer "long tail." Even if the platform does not delete the post, a single misinterpretation or screenshot forwarding by an acquaintance can have long-term relationship consequences.

One participant explained:

> *"If the platform deletes it, you can pretend it never happened, but if an acquaintance screenshots it and sends it to others, you'll be explaining yourself for a lifetime."* (P7)

Crucially, this interpersonal risk extends beyond mere social awkwardness into what Andrejevic [90] describes as "lateral surveillance"—a condition where peer monitoring reinforces institutional control. In this specific context, the "acquaintance" is not just a social connection but a potential proxy for systemic risk. Participants expressed a distinct fear of the "weaponization of context collapse," where platform affordances like screenshots allow private speech to be decontextualized and recirculated as evidence for reporting. Unlike the automated and often impersonal nature of algorithmic censorship, this risk comes from human actors who understand the hidden transcripts and can decode the irony that algorithms might miss. Therefore, the "snitch" in the friend circle is perceived as a more unpredictable threat than the censor in the machine.

Interestingly, many participants believe that interpersonal risk is more difficult to control. While systemic risk is opaque, at least "the platform is an object"; however, the risk in close-knit networks stems from the relationship itself, and intimate relationships imply more real-world consequences, with misunderstandings being harder to correct. This also explains why many participants tend to use stronger denial and more ambiguous expressions in familiar settings, or even choose silence or migrate to semi-anonymous spaces.

4.4.2 "Scene safety" is dynamically assessed: observing the wind direction—choosing a location—timely retreat

Participants rarely view platforms as "always safe" or "always dangerous." They are more like conducting a continuous, dynamic assessment: observing the wind direction before speaking out, choosing a location to express themselves when speaking out, and quickly retreating or migrating when the risk increases. This dynamic assessment permeates the specific operational level. For example, one participant explained:

> *"You know, power permeates everywhere (on social media platforms), no place is absolutely safe, but it's not that dangerous either. Believe me, freedom cannot be destroyed or killed. No matter how strict the prohibitions, no matter how terrifying the tyrant, no one has the brains to completely suppress all freedom."*

The first step is "observing the wind direction." Participants will first see if similar content appears in large quantities, if the atmosphere in the comment section is tense, if anyone mentions being "deleted/limited/warned," and even observe whether "pinned comments" and "hot comments" have been replaced. Judging the direction of public opinion does not require explicit evidence; it often relies on subtle signals: a sudden

decrease in comments (P1), unusual collapsed notifications (P8), or people starting to use more "you know what I mean" or "..." to soften the tone (P19).

One participant said:

> *"I'll first refresh the page to see if anyone says 'stop talking' or 'be careful.' If the comment section starts collectively adding dog emojis and quotation marks, I know something's not right today."* (P19)

Secondly, there is "choosing where to express oneself." Participants will place the same viewpoint in different locations with varying intensity: the main post is often more prominent and easier for the system or strangers to pick up; the comment section is more fragmented and easier for collaborative repetition and meaning completion (corresponding to the sixth type of relay mechanism); smaller group chats/private chats, while having lower systemic risk, may have higher interpersonal risk, thus requiring more careful audience selection.

One participant explained:

> *"I'm not too sure about explaining things clearly in the main post, so I'll just leave a comment like 'the familiar formula,' and those who understand will understand; if I really want to explain, I'll do it in a smaller group."* (P11)

Finally, there is "timely retreat": when participants sense increased risk, they choose to delete, edit, move to a secondary account, or simply stop discussing. Many people treat "leaving a message before deleting" as a ritualistic signal: it is both a reminder to others that "this post might not last long," and a public declaration of the expresser's risk expectation.

One participant described this as:

> *"Sometimes I take a screenshot after posting, assuming it won't last long. Posting 'leaving a message before deleting' is essentially telling everyone: don't elaborate here."* (P14)

These dynamic strategies make "safety" a process rather than a state: the safety of a scenario is not a platform attribute, but rather the result of changes in platform distribution, discussion intensity, and relationship networks.

4.4.3 Six Types of Self-Protection Strategies: From "Tone Buffering" to the Order of Escalation of "Location Migration"

In terms of specific strategies, the practices described by participants show a clear hierarchical structure: they do not use the most extreme migration or silence from the beginning, but rather gradually "escalate" the measures. This escalation order highly corresponds to the types of jokes mentioned earlier—the higher the risk, the more the expression tends towards the second, fourth, and third categories; when the risk is lower, the first and seventh categories can be fully developed.

The first strategy is pragmatic buffering: using facial expressions, interjections, exaggerated tone, or "haha/cracking" to disguise the expression as a joke. This strategy is closely related to the fourth type of tone buffering and the first type of semantic reversal—"dog head," "quotation marks," and "..." do not change the core semantic meaning, but they change the pragmatic stance, reducing aggression and certainty.

"I'll see if I can use 'laughing to death' or 'dog head' to gloss over it, making it seem like a joke, so others won't take it too seriously." (P25)

The second strategy is outsourcing responsibility: using "retelling," "it is said," "I heard from a friend," or "this is only my personal opinion" to change the stance from "I said" to "I am just quoting." This corresponds to the second type of denial stance, the core of which is to blur the subject of accountability.

"Saying 'it is said' is different; you're relaying information, not drawing conclusions. If questioned, you can backtrack." (P26)

The third strategy is semantic encryption: misspellings, homophones, abbreviations, white space, and censorship make the text difficult to directly retrieve, while requiring insiders to fill in the meaning through context. This, along with the fourth type of tone buffering and the third type of referential shift, constitutes a "language cue package," a key practice in the RQ2 coding stage mentioned earlier.

"I'll write the keywords slightly off-key, or simply use pinyin abbreviations. Those who understand will immediately grasp it; those who don't will see it as gibberish." (P3)

The fourth strategy is audience segmentation: the same sentence is expressed with different intensities to different audiences, or the context is changed entirely. Participants break the content down into three layers: the public area only leaves one or two debatable hints; the comment section uses collaborative repetition and relay completion; and the truly specific discussions are moved to small groups or private chats. This segmentation is driven by the fear of lateral surveillance mentioned above; participants actively filter out "untrustworthy acquaintances" who might weaponize the content.

"In public, I'll just say a few words and see if anyone responds in the comments; if you really want to go into details, go to a small group, otherwise you're just asking for trouble." (P4)

The fifth strategy is media substitution: using memes, screenshots, and emojis to replace searchable text expressions. Participants believe that images, on the one hand, reduce the searchability of text, and on the other hand, better fit the disguise logic of "meme play," making the expression seem like entertainment rather than political discussion.

"Text is too direct; I'd rather post pictures. With pictures, you can say 'I just think it's funny.'" (P15)

The sixth strategy is time and identity strategy: deleting posts after the hype has died down, using alternate accounts, only posting on short-lived channels, or even choosing silence. This level reflects that when the risk cannot be resolved through rhetoric, expression will turn to a structural withdrawal.

"If it doesn't work, delete it, or change your number. Ultimately, that's all you can do; in the end, you just don't say anything." (P11)

Many participants explicitly described an escalation path (20 out of 26): when the risk escalates, they first "add buffers and escape routes" within the text (tone buffering, outsourcing responsibility). If they are still uneasy, they then change visibility and searchability through "encryption and migration" (semantic encryption, audience segmentation, media replacement), ultimately choosing silence or retreat in the most uncertain situation.

> *"I usually change the tone first, then write the words a little crookedly. If that doesn't work, I'll say it in a group chat; if I'm still uneasy, then I'll just give up."* (P12)

4.5 RQ4: The Consequences of Joking – Safe Resonance and Self Reshaping

In the questionnaire design, we explicitly asked participants to recall both experiences that "worked very well" and experiences that "caused problems," and to inquire about their long-term impact. Interview results show that the consequences of joking have a clear double-edged nature: it provides a low-risk outlet for expression, but it can also have lasting effects on relational and psychological levels. More importantly, these consequences are not the end point, but rather reshape an individual's future expression strategies and risk thresholds.

4.5.1 Positive Consequences: Low-Risk Resonance, Community Membership, and "Being Seen"

When others successfully responded to a joke, repeated it, or escalated the conversation, participants generally described a "safe resonance" (mentioned by all 26 participants). This resonance did not stem from explicit positional debate, but from the confirmation of a shared context. For example, when a "familiar recipe" is quickly followed by "History always repeats itself," "+1," and "Let's form a group," the speaker feels that "the code has been matched." This type of resonance does not require explaining a position or elaborating on arguments, yet it confirms mutual understanding of reality under low-risk conditions. Participants pointed out that this "being seen" does not come from the number of likes, but from being "correctly interpreted."

> *"It's not about how many people like it, but that someone responds particularly accurately, and you know you're 'on the same wavelength.'"* (P1)

This effect reinforces community membership. The first type of semantic reversal joke (such as "We really won again") and the third type of referential shift (such as "a certain major power" and "across the sea") form stable identification markers within the community, while the sixth type of collaborative repetition further consolidates this group identity. Expression thus becomes a low-cost mechanism for maintaining community. It makes individuals feel less isolated and provides emotional support for sustained attention to public issues.

Mechanistically, the positive consequence of banter is that it transforms high-risk expressions of stance into low-intensity shared signals, maintaining a minimum level of connection in public discussion even under pressure.

4.5.2 Negative Consequence 1: Misunderstandings and Conflicts—Especially in Cross-Circle/Acquaintance Contexts

However, banter is highly dependent on contextual consensus. Once outside the original circle, its two-way readability can quickly fail. Participants recalled that most "incidents" were not due to direct platform penalties, but rather to cross-circle misunderstandings.

The first scenario is that irony is taken seriously. For example, phrases like "open-mindedness" and "gratitude" are seen as genuine support, not irony, and the speaker is labeled with a stance.

> *"I meant it sarcastically, but a relative actually thought I was praising someone and even replied in the group chat, 'You finally figured it out.'"* (P12)

The second scenario is being accused of being "sarcastic." Some acquaintances hold a morally critical attitude towards sarcastic tones, believing that such expressions are "not positive enough" and "not responsible enough."

> *"Some people will say, 'Just say what you mean, don't be sarcastic.' But saying it directly is the most dangerous thing."* (P13)

The third scenario is being screenshotted and spread. Especially among acquaintances, even if an expression does not trigger systemic risks, it can still be forwarded to a wider network of relationships, causing a long-tail effect. Participants described this consequence as "extremely costly to explain."

> *"Deleting posts is useless; others have already screenshotted them. You have to explain to each person individually, which is more tiring in the end."* (P15)

Therefore, many people are gradually withdrawing political satire from their WeChat Moments to smaller circles or moving it to anonymous spaces. This "expression migration" is highly related to the audience structure reasoning mentioned earlier. The risk is not just a platform issue, but also a relationship issue.

4.5.3 Negative Consequence Two: The Accumulation of Political Cynicism and Self-Nutrition

Besides immediate conflict, participants also reflected on the impact of satire on long-term mentality. Some realized that while frequent use of the first type of irony and the seventh type of humorous expression reduced emotional conflict, it could also weaken serious discussion of the issues.

> *"At first, it was to avoid a breakdown, but then I found myself increasingly too lazy to speak seriously, since it all ended up becoming a meme anyway."* (P1)

The fifth type of cynical expression (such as "It's all the same," "Don't take it too seriously," "There's nothing new under the sun") is gradually shifting from a one-off response to a normalized stance. This change echoes the long-term psychological defenses mentioned earlier. Expression is no longer an emotional adjustment to specific events, but a strategy to cool down the overall public environment.

Mechanistically, banter protects individuals in the short term, but in the long term, it may change the form of public discussion: serious issues are constantly turned into jokes, and discussions shift from argumentative to gestalt-like. This chronic effect is not necessarily the result of the expresser's active choice, but rather an adaptive byproduct of a risky environment.

4.5.4 Feedback Effect: A "Disaster Experience" Can Rewrite Expression Strategies

Participants generally mentioned that a negative experience often becomes a turning point. Whether it's being deleted, having their reach limited, being reminded, or being misinterpreted or screenshotted by an acquaintance, it can change their subsequent expression methods.

> *"After being reminded that time, I basically stopped posting in the main thread and just left a comment."* (P4)

> *"After being misunderstood by a relative, I almost never posted anything related to that on my WeChat Moments."* (P7)

These experiences not only changed specific sentence structures but also altered the expression location, audience selection, and risk threshold. Some people started using the second type of denial more frequently, such as:

> *"It's better to express things in a safer way possible."* (P4)

Some strengthened the fourth type of tone buffer, such as:

> *"You can't be too sure of what you're saying, even when it's joking."* (P11)

Some turned to the sixth type of collaborative repetition to avoid original expression, such as:

> *"Repetition is always safe."* (P22)

And some migrated directly to anonymous platforms or chose silence, such as:

> *"I don't use WeChat to joke anymore. On Bilibili, I won't be recognized by acquaintances; it's a stranger-based social network."* (P17)

This shows that the consequences are not the end of the expression cycle but rather the input conditions for the next round of expression decisions. The practice of joking presents a feedback loop: Motivation → Encoding → Collaborative Relay → Consequences → Risk Threshold Adjustment → Next Round of Expression. An "incident" may reduce the intensity of future expression or lead to more covert, decentralized and decentralized forms of expression.

5 Discussion

5.1 Appropriate Amount of Publicity under Content Moderation: Unique Narrative Strategies for Political Discourse

The euphemistic strategies adopted by users when joking about politics raises the cost of censorship by creating detection barriers for manual and machine-learning-based filtering systems [8, 11, 51]. These strategies are not adopted for maximizing publicity, as posts become more visible, they become easier to spot for review and can trigger faster removal. Instead, as research suggests, participants described optimizing for a more controlled visibility, remaining socially present while keeping meanings selectively legible to the "right" audience and plausibly dismissible to outsiders [91]. Participants repeatedly emphasized that greater publicity can itself become the trigger for enforcement, so sacrificing exposure range through euphemism, "inside" references, and joke-like packaging is not a fallback but a deliberate choice. This aligns with how people manage collapsed audiences online: when the same utterance can be encountered by strangers, acquaintances, and potential reporters simultaneously, the safest form of publicity is often one that is selectively readable to intended audiences while plausibly dismissible to others [92, 93]. Under such conditions, "appropriate publicity" is achieved not by broadcasting more clearly, but by engineering ambiguity in ways that still allow recognition among those who share context. This "right-audience readability" also connects to how trust and credibility judgments can depend on social ties and crowd interpretation in regulated media environments, making the comment layer an expressive resource [53].

Although these strategies can provide a shield against censorship, users adopting different strategies can pose uneven levels of difficulty for systematic detection. For exaggerated irony (e.g., "They really won again.") or even more parodic satire ("Whether or not you believe it, I believe it" from Lee [81]) that mimics authoritative reasoning, the surface form can remain formally compliant while the critique is displaced into contextual inference. This is difficult to detect without overreach: the literal text may appear positive or neutral, and the "political" meaning emerges only through shared background knowledge and the recognition of semantic tension. In comparison, tone buffering often relies on relatively surface-level stance markers (e.g., quotation marks, emojis) that can be more straightforward to enumerate or model than context-dependent irony, even if the intended meaning still depends on shared inference [94]. Users may voluntarily or involuntarily choose strategies with different levels of safeguard depending on their sociotechnical background and conception of censorship [55], and by continually reading the platform "wind" updating expectations based on how the feed feels, what survives, and what gets corrected or removed.

Taken together, our results suggest that "appropriate publicity" is not just a compromise between speaking and silence, but an ongoing calibration problem: users combine narrative strategies to keep messages socially readable, collaboratively interpretable, and systemically debatable at the same time. Rather than treating euphemism and jest as lower-quality substitutes for direct speech, we show how they become a rational, compositional toolkit for distributing risk across language (how

to say), interaction (how others relay), and audience boundaries (where meaning is completed).

5.2 Multilateral Infrastructure: The Inner and Outer Circles' Interaction on Collaborative Expression

Our findings portray censorship response as a multi-player governance problem rather than a simple battle between "platform" and "user". The outer circle includes platform moderation and state-linked enforcement pressures, reflecting how private platforms can become part of broader regimes of networked control. These forces form an external censorship infrastructure that is pervasively embedded into daily social life through networked information controls (filtering, takedowns, and surveillance-enabled enforcement). The existence of censorship infrastructure is widely acknowledged, felt, and studied [54, 95]. Participants stressed that moderation triggers are usually unknown and ex post facto [11, 46]. Therefore, they develop folk theories [61, 62, 71, 96] to collectively understand the mechanism of moderation, and then adapt iteratively by developing counter-mechanism for such moderation - learning from sanctions, warnings, deletion, or subtle signals in the feed and comment atmosphere. They can thus maintain the long-term stability on platforms by carefully managing the cost of misinterpretation and leveraging the mutually acceptable intensity of expression ("Those who know, know") reducing the systemic risk.

At the same time, the inner circle, friends, acquaintances, and regular audiences, also functions as an infrastructure of constraint. [97] The reliance on relational infrastructure creates distinctive vulnerabilities. For users adopting such political joking mechanism, they also have to face the lack of understanding and misunderstanding from users who are accustomed to the moderation or censorship [53]. Irony can be taken literally and lead to unwanted labeling; conversely, sarcasm can be morally condemned by those accustomed to "proper" speech or who interpret joking as irresponsible. These dynamics mean that users sometimes deploy stronger deniability in familiar settings [92]. The interpersonal risk including misunderstanding, moral judgment about sarcastic tone, and screenshot circulation within acquaintance networks, was described as having a longer and more exhausting tail than content deletion.

To face both inner and outer circles' systematic infrastructure on moderation and censorship, the expressive engineering mechanism in fact is also becoming infrastructural or at least systematic, requiring pervasive collaboration and designed strategies. Thus, the platform and users co-evolve: the platform tolerate mild expression such as jest and satire as the expressivity is the key factor to maintain user engagement and vivid ecosystems [98], but censor or moderate the narratives over the borderline out of the requirement of regulation compliance; users will continue to contribute to the vivid expressive discourse online to interact with others for publicity, and learn to use their "folk theories" to make sense of the censorship infrastructure from their censored discourse because of infraction, and reinforce engineering on their collaborative expression, and stay in the tolerant zone of the platform. Therefore, the platform and users are not antagonistic, they are both evolving component of the information ecosystem under the censorship.

5.3 Normalization and Internalization

The feedback loop we identify implies that repeated participation in the collaborative political expression does more than help users familiarize censorship in the short term; it can reshape how political talk is experienced and what forms of participation feel normal("At first, it was to avoid a breakdown, but then I found myself increasingly too lazy to speak seriously, since it all ended up becoming a meme anyway."). Our finding that the exposure to censorship and settle down adopting resilient strategies to cope with censorship clearly aligns with the study that the usage of information technologies will accommodate or "normalize" users [55, 99]. From this perspective, euphemistic collaboration can look like resilience against censorship while also revealing accommodation: users increasingly treat constraint as a normal condition of participation, and they internalize it by reorganizing their daily expressive practices around risk.

On the positive side, participants described safe resonance as a low-risk way to meet a basic need for expression. They could still signal stance, find people who "know" and feel less alone because the message reached the right readers who could interpret it correctly, realizing collaborative relay. In this sense, joking sustains participation under constraint: people keep speaking (and keep each other speaking) by treating censorship as something to be managed in everyday practice rather than confronted head-on [45].

But this safety comes with a significant cost. The lowered risk is achieved indirectly through ambiguity, detours, and self-limiting routines rather than through freer speech. Over time, this indirectness can normalize censorship as the "normal" environment of public talk: users learn what survives, and then design their expression around survival instead of clarity. This is exactly why the tradeoff is so consequential: the system does not only remove content; it reshapes what people consider safe to say and how much effort they are willing to invest in saying it [55, 100].

Participants also pointed to a deeper internalization effect: once joking becomes the safest default, people may start treating more direct engagement as unrealistic because the cost-benefit calculation repeatedly favors safer, more indirect forms. Participants' broader information environment can reinforce these dynamics: censorship awareness intertwines with perceptions of misinformation and manipulation, which can increase confusion and indirectly strengthen pro-control attitudes [46], while the co-existence of censorship with influence operations can further reconfigure what users perceive as safe, feasible agency in everyday participation [77]. In the long run, this is one pathway into resignation: people adapt to unwanted sociotechnical conditions because meaningful control feels out of reach [101].

Such chilling dynamics can weaken public discourse: regulatory or surveillance pressure can reduce information seeking and discourage open participation, which is corrosive to healthy political discussion. Meanwhile, platforms face a structural tension: they often present themselves as open spaces for participation and community, yet must impose rules and enforcement that narrow what can be said and seen [14]. As private governors of online speech, platforms curate and regulate expression in ways that can leave users with less power to monitor governance and execute their rights [98].

6 Limitation and Future Work

This study has several limitations that also suggest directions for future research.

First, the study is based on semi structured interviews. Participants described satirical expressions from memory. This approach helps us understand their motivations, interpretations, and risk considerations. However, we did not directly observe the full interaction on the platforms. Important dynamics such as how comments unfold, how users respond to each other, and how expressions change over time cannot be fully reconstructed. Future work could analyze real comment threads or platform data with computational approaches to better understand how satirical expressions spread and develop in online discussions, providing a more comprehensive understanding of collaborative expression dynamics.

Second, the study focuses on Chinese social media platforms. Political expression in this environment is shaped by both platform moderation and state censorship. These conditions influence how users assess risks and design their expressions. Some patterns we identified may also appear in other regions and platforms, but the specific dynamics may differ across political and cultural contexts. Future research could investigate how expressive mechanisms vary across different platform environments. Differences in recommendation algorithms, comment infrastructures, and moderation practices may significantly shape how satirical expressions are produced and disseminated. Comparative studies across platforms could therefore help illuminate how platform infrastructures influence expressive strategies and interaction patterns.

Finally, this study examines about one hundred satirical examples shared by participants statically. The goal is not to list every form of political satire, but to understand how these expressions are created and interpreted through interaction. Applying longitudinal data to study such interaction will provide a larger landscape of affected range and a thorough manifestation of satirical expression lifecycle. Future research could further examine the long term effects of satirical expression, including how it shapes political discussion, community relationships, and participation in public issues over time.

7 Conclusion

Our findings suggest that political satire in highly moderated environments should not be understood as a simple evasion tactic. Instead, it functions as a practical system for calibrating *appropriate publicity*. Users remain public enough to signal social presence, while keeping their expressions ambiguous enough to avoid concentrated punishment.

Across interviews, participants described several recurring expressive practices, including semantic reversal, denial posture, referential shift, tone buffering, collaborative repetition, and anecdotal narration. These practices reflect motivations such as risk avoidance, emotion regulation, and the need to remain socially present in public discussion.

We show that satirical expression emerges through an interactional cycle of Trigger, Encoding, Relay, and Stabilization. Meaning is first encoded in the initial post. It is then expanded through comment based interaction. Over time, users adjust their

expressions in response to consequences such as deletion, misunderstanding, screenshot circulation, or successful resonance.

8 AI Use Statement

Generative AI tools were used solely for language refinement and minor structural editing during manuscript preparation. No AI system was involved in data analysis, coding, theoretical development, or interpretation of findings. All research decisions, analytical frameworks, and conclusions are the sole work of the authors. The authors remain fully responsible for the content of this manuscript.

9 Positionality and Political Neutrality Statement

This study does not seek to endorse, oppose, or evaluate any specific political actors, institutions, or ideological positions. Instead, it examines communicative practices and collaborative expression mechanisms as sociotechnical phenomena within particular digital environments. The analysis focuses on how users navigate perceived risks, audience structures, and platform affordances when engaging in public discourse. The authors approach the topic from an interactional and analytical perspective, aiming to understand pragmatic strategies rather than to make normative claims about political systems or policies.

Acknowledgements. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors conducted this research independently and received no external financial support for the design, execution, analysis, or publication of this study.

Appendix

1. Participant Demographics and Social Platform Use

ID	Age	Sex	Occupation	Highest degree	Urban/Rural	Use social platforms
1	19	M	Student	High School	U	Bilibili, Tiktok, Zhihu, Tieba, Wechat, Weibo
2	21	F	Farmer	Primary school	U	Bilibili, Rednote, Tiktok, Wechat, Weibo
3	22	F	Student	Bachelor's Degree	U	Bilibili, Rednote, Zhihu, Tieba, Wechat, Weibo
4	23	M	Writer	Bachelor's Degree	U	Bilibili, Rednote, Zhihu, Tieba, Wechat, Weibo
5	24	M	Teacher	Master	U	Bilibili, Rednote, Tiktok, Zhihu
6	25	M	Chef	High School	R	Bilibili, Rednote, Tiktok, Tieba, Weibo
7	27	F	Editor	PhD	U	Bilibili, Rednote, Tiktok, Zhihu, Tieba, Wechat, Weibo
8	30	F	Labour	Junior high school	R	Bilibili, Rednote, Wechat, Weibo
9	29	M	Programmer	Bachelor's Degree	R	Rednote, Tiktok, Zhihu, Tieba, Wechat, Weibo
10	24	M	Barista	Primary school	R	Bilibili, Tiktok, Zhihu, Tieba, Wechat, Weibo
11	25	F	Baker	Bachelor's Degree	U	Bilibili, Tiktok, Zhihu, Tieba, Weibo
12	27	F	Designer	Bachelor's Degree	R	Bilibili, Rednote, Tiktok, Zhihu, Wechat
13	22	M	Student	High School	U	Bilibili, Rednote, Tiktok, Tieba, Wechat, Weibo
14	24	M	Data Analyst	Bachelor's Degree	R	Bilibili, Rednote, Tiktok, Zhihu, Tieba, Wechat, Weibo
15	18	F	Student	High School	U	Bilibili, Rednote, Tiktok, Tieba, Wechat
16	19	F	Student	High School	R	Bilibili, Rednote, Tiktok, Zhihu, Tieba, Wechat, Weibo
17	19	M	Labour	High School	R	Rednote, Tiktok, Zhihu, Tieba, Wechat, Weibo
18	19	F	Student	High School	U	Bilibili, Rednote, Tiktok, Zhihu, Tieba, Wechat, Weibo
19	26	F	Manager	High School	U	Bilibili, Rednote, Tiktok, Zhihu, Tieba, Weibo
20	27	M	Journalist	Bachelor's Degree	R	Bilibili, Zhihu, Tieba, Wechat
21	29	M	Accountant	Bachelor's Degree	R	Bilibili, Rednote, Tiktok, Tieba, Wechat, Weibo
22	30	M	Photographer	Primary school	U	Tiktok, Zhihu, Tieba, Wechat
23	26	F	Designer	Bachelor's Degree	U	Bilibili, Rednote, Tiktok, Tieba, Wechat, Weibo
24	26	F	Photographer	Bachelor's Degree	U	Bilibili, Rednote
25	27	M	Doctor	Bachelor's Degree	R	Bilibili, Tiktok, Tieba, Wechat
26	29	M	Programmer	Bachelor's Degree	R	Bilibili, Tiktok, Tieba

Table 1 Participant demographics and social platform use.

2. Semi-Structured Interview Guide

Opening and Safety Statement (1–2 minutes)

"Hello. In the following conversation, we will talk about some satirical or humorous expressions related to public issues. You may paraphrase, rewrite, or censor details. There is no need to mention specific names, organizations, or events. We are interested in the form of expression and the process of using it."

Platform and Usage Background (about 5 minutes)

- "What social media platforms do you use most frequently? Please describe them in order of usage frequency. On these platforms, do you mainly browse, like, comment, repost, or create original content?"
- "Do you more often speak in publicly visible areas, or in spaces visible to acquaintances such as Moments, group chats, or private chats? Please give one example of your most common scenario for each."
- "In the past six months, how often have you discussed public issues? Has it been more or less than before? What has changed?"

1). Collection and Interpretation of Satirical Expressions (about 20 minutes)

Moderator Instruction

- The goal of this section is to let participants provide as many expressions as possible (at least four), but only give a "light interpretation" for each one without going too deep.
- Then the participant selects four expressions as focal examples (S1–S4). All subsequent modules will revolve around these four.
- "Please mention as many satirical or humorous expressions related to public issues as you have seen or used in the past six months. You may quote them directly or paraphrase. You do not need to describe specific events—just explain what kind of discussion context they usually appear in."

Follow-up prompts (if participant gets stuck)

- "You can think about: common short comments in comment sections, patterned repetition, ironic statements, historical references, feigned ignorance, use of dog-head emojis/quotation marks/ellipses, homophones or deliberate misspellings, meme captions."

Quick Interpretation Card for Each Expression (30–60 seconds per item)

For each expression mentioned, ask the following three questions:

- "What is the literal meaning of this sentence?"
- "In that context, what do people usually understand it to actually mean?"
- "What do you think mainly makes it understandable to those who 'get it'? For example, irony, insinuation, feigned ignorance, symbols,

patterned repetition, meme format... You can describe it in your own words."

Select Four Focal Expressions

- "Among the expressions you just mentioned, please choose four that you think are most typical, most frequently seen, or best represent different usages. We will focus on these four next."

The moderator writes down the four expressions and labels them S1–S4.
If the selected four are very similar, the moderator asks:

- "Could you choose one that is clearly different? For example, one that is more like patterned repetition, one more like irony, one more like feigned ignorance, one more like historical reference or meme format."

2). Why Use These Satirical Expressions (about 15 minutes)

Overall Motivations

- "Overall, why do you think people use this kind of satirical expression instead of speaking directly? You can start from your own experience."
- "For you personally, what are the most common motivations? For example: to avoid risk, to regulate emotions, to find like-minded people or signal presence, to create distance and make things more bearable, to feel safer through group participation, or gradually becoming cynical or numb. You do not have to use my terms—please describe in your own words."
- "If you had to rank them, what are the two most important motivations for you? Why?"

Align Motivations with S1–S4 (2–3 minutes each)

- "Taking S1 as an example: what motivation does it mainly serve? If you could only choose one core motivation, which would it be? Why?"
- "If you did not use satire but spoke directly instead, what do you think would happen? Would it be more effective, or more dangerous, or more likely to lead to conflict?"

3). How They Are Used: Creation, Circulation, and Collaborative Understanding (about 20 minutes)

- "For an expression like S1, do you remember under what circumstances it usually appears? For example, when a hot event just breaks out and information is chaotic, or when a certain narrative template reappears?"
- "Have you seen cases where one comment suddenly 'hits the point,' and then many people start copying it? Can you describe that process?"
- "When you write or repost such satire, do you deliberately try to make it understandable to those who 'get it,' while unproblematic to those who do not? How do you usually achieve that?"
- "Do you add 'exit routes' in your expression, such as pretending to quote someone else, pretending not to understand, adding emojis/quotation marks/ellipses, or misspelling key words? Can you describe a memorable example?"
- "Do you think the meaning of these satirical expressions is often not completed in one sentence, but gradually completed in the comment section through repetition, chaining, upgrading, or adding another line? Can you describe a time when you saw people 'each add one sentence'?"
- "When someone speaks too directly, have you seen others remind or correct them? For example, asking them to delete a word, add a dog-head emoji, or rephrase. What do you think such reminders are doing?"
- "Have you seen satire being taken literally or misunderstood? When misinterpretation occurs, how do people usually handle it? Do they clarify explicitly, or shift to a more ambiguous explanation? Why?"
- "When someone directly asks you 'What do you mean?', how do you usually respond? Do you explain, pretend not to understand, or change the topic? Can you give a personal example?"

4). Safety and Risk (about 15 minutes)

- "When you are about to post satire like S1–S4, does the thought 'this might be inappropriate or have consequences' come to mind? What is the first consequence you usually worry about?"
- "Are you more concerned about platform-level consequences (deletion/visibility reduction/warnings/account suspension), or interpersonal consequences (misunderstanding by acquaintances/screenshots/relationship impact)? Why?"
- "How do you judge whether a scenario is safe or not? What signals do you observe? For example, comment section atmosphere,

whether someone says 'be careful,' changes in top comments, folding prompts, or people suddenly collectively adding quotation marks or dog-head emojis?"

- "When you feel risk increasing, how do you gradually escalate your protective strategies? Please describe in order: how you revise before posting, how you write during posting, and how you handle it afterward."
- "Have you ever experienced a situation where you thought it was safe but it turned out not to be? After that, what changed? Was your 'safety threshold' rewritten?"

5). Consequences (about 15 minutes)

- "After using satire like S1–S4, what do you observe to judge whether it achieved its purpose? Do you care more about the number of likes, or about others picking up the joke, repeating it, upgrading it, messaging you privately, or misunderstanding it?"
- "Describe one experience where you felt it 'worked very well': others responded precisely, and you felt understood or seen. What impact did that have on you?"
- "Describe one experience where it 'went wrong': misunderstood, attacked, questioned by acquaintances, screenshot and circulated, deleted/visibility reduced/warned. How did you handle it at the time? Did you change your expression style afterward?"
- "In the long term, what effect does this kind of satire have on your emotions? More stress relief, anxiety, numbness, anger, or a kind of chronic defense?"
- "Overall, do you think this kind of satire makes public discussion more active, more implicit, or more cynical? Why? Have you yourself therefore expressed directly less often, and shifted more toward small groups or silence?"

Closing (2–3 minutes)

- "If you had to summarize in one sentence: in your view, what is the biggest function of these satirical expressions in a constrained environment?"
- "Do you think I missed any key point? Is there something we must understand that I did not ask about?"

3. Political Satirical Expressions Collected in Interviews

No.	Original Expression (Chinese)	English Rendering	Description
1	真是又赢了	They really won again	Used sarcastically to question official narratives portraying success.
2	赢麻了	Winning too much	Used ironically to criticize exaggerated narratives of achievement.
3	我们中国太厉害了（鼓掌）	China is amazing (applause)	Used ironically to question official praise of national achievements.
4	咱就是说，这波操作属实高端	This move is extremely sophisticated	Used sarcastically to criticize political decisions.
5	格局打开	Open up the perspective	Used ironically when negative outcomes are reframed positively.
6	遥遥领先	Far ahead	Used sarcastically to question narratives of superiority.
7	感恩	Grateful	Used ironically to criticize official discourse.
8	太感动了	So touching	Used sarcastically to question ideological narratives.
9	建议申遗	Suggest applying for UNESCO World Heritage status	Used to mock absurd or unethical behavior.
10	这就是实力，蒸馍，泥不福	This is real strength	Used sarcastically to question official narratives.
11	这波是虽败犹荣	A glorious defeat	Used sarcastically when failure is framed as success.
12	这也是一种成功	This is also a kind of success	Used ironically to reinterpret negative outcomes.
13	战略性调整	A strategic adjustment	Used sarcastically to describe decline framed as policy success.
14	高质量人口增长	This is good news for population quality	Used sarcastically to discuss declining fertility rates.
15	长远来看是好事	In the long run this is good	Used ironically when a situation is clearly negative.
16	我们至少...	At least we...	Used to mock attempts to frame poor outcomes positively.
17	换个角度看	Look at it from another angle	Used sarcastically to reinterpret negative situations.
18	格局要大	Think big	Used ironically when individuals are asked to tolerate losses.
19	精神胜利法	Spiritual victory	Used sarcastically to describe rationalizing failure.
20	赢在未来	We will win in the future	Used ironically to postpone recognition of failure.

No.	Original Expression (Chinese)	English Rendering	Description
21	节目效果拉满	The show effect is maxed out	Used to mock performative political statements.
22	年度喜剧	Comedy of the year	Used sarcastically to describe absurd discourse.
23	笑死	That's hilarious	Used dismissively to reject a viewpoint.
24	太真实了	That's so real	Used to express agreement with non-mainstream views.
25	活久见	You live long enough to see this	Used to express disbelief toward events.
26	我不懂但我大受震撼	I don't understand but I'm shocked	Used to signal disagreement while claiming ignorance.
27	我只是路过	I'm just passing by	Used to distance oneself from sensitive content.
28	我鹦鹉丢了，本人不同意本人鹦鹉的观点	My parrot said it, not me	Used to outsource responsibility for opinions.
29	求科普	Could someone explain	Used to indirectly question sensitive issues.
30	听不懂思密达	I can't understand this	Used to avoid expressing a clear stance.
31	真的吗?	Really?	Used rhetorically to signal doubt.
32	我不好评价	I can't comment	Used to imply criticism without explicit evaluation.
33	咱也不敢说	I dare not say	Used to signal restrictions on expression.
34	你说得都对	You're absolutely right	Used sarcastically to express disagreement.
35	行吧就当我没说	Pretend I didn't say anything	Used to retract statements.
36	别问我	Don't ask me	Used to emphasize expression constraints.
37	仅代表个人观点	This is only my personal opinion	Used to weaken responsibility.
38	不针对任何人	Not directed at anyone	Used to soften criticism.
39	如有雷同纯属巧合	Any resemblance is coincidental	Used to indirectly reference real actors.
40	仅供娱乐	For entertainment purposes only	Used to frame political speech as humor.
41	不要上纲上线	Don't politicize it	Used to discourage political interpretation.
42	请勿过度解读	Please don't overinterpret	Used to reduce semantic certainty.
43	我只是转述	I'm just relaying what I heard	Used to shift responsibility.

No.	Original Expression (Chinese)	English Rendering	Description
44	据说	It is said that	Used to frame statements as rumors.
45	听朋友讲	A friend told me	Used to attribute information to a third party.
46	删前留念	Screenshot before deletion	Used to anticipate censorship.
47	颇有蒋匪遗风	Reminds me of Chiang's regime	Historical metaphor criticizing contemporary politics.
48	熟悉的配方	The familiar formula	Used to suggest recurring patterns.
49	祖传四代庸医	Four generations of quack doctors	Used metaphorically to criticize institutional continuity.
50	历史总是惊人相似	History always repeats itself	Used to criticize events through historical analogy.
51	大清还没亡	Has the Qing dynasty not fallen yet	Used sarcastically to criticize authoritarian practices.
52	后清	The later Qing dynasty	Used metaphorically to criticize governance.
53	穿越回去了	It feels like going back in time	Used to describe perceived regression.
54	山顶洞时代	Back to the cave era	Used hyperbolically to criticize decline.
55	又到了考古时间	Time for archaeology again	Used to mock revival of outdated ideas.
56	修史了	History is being rewritten	Used to criticize narrative revision.
57	1984不是操作指南	1984 is not an instruction manual	Literary reference criticizing surveillance.
58	北方的朋友	Our friends in the north	Vague geographic reference.
59	岛那边	Things are quiet on the island	Indirect geographic reference.
60	隔壁国家	The neighboring country	Indirect reference to another state.
61	某大国	A certain major power	Indirect reference to a country.
62	宇宙第一	Number one in the universe	Used sarcastically to mock nationalism.
63	地球第一	Number one on Earth	Used sarcastically to criticize superiority narratives.
64	世界灯塔	The lighthouse of the world	Used ironically to critique moral self-positioning.
65	东方神秘力量	The mysterious power of the East	Used humorously to explain phenomena.
66	海对面	Across the ocean	Indirect reference to the US.

No.	Original Expression (Chinese)	English Rendering	Description
67	东方大国	The great Eastern power	Euphemistic reference to China.
68	懂的都懂	Those who understand will understand	Used to imply shared understanding.
69	.....	...	Silence used to signal dissatisfaction.
70	呵呵	Hehe	Used sarcastically to express disagreement.
71	笑而不语	Smile without speaking	Used to imply silent criticism.
72	狗头	Dog emoji	Used to signal joking or irony.
73	括号保命	Parentheses for safety	Used to reduce perceived risk.
74	打个引号	Put it in quotation marks	Used to question authenticity.
75	我打错字了	I made a typo	Used to pretend expression was accidental.
76	拼音缩写	Pinyin abbreviation	Used to evade keyword filtering.
77	消音	Name censored	Used to avoid naming actors.
78	666	666	Used sarcastically to signal mock admiration.
79	早就习惯了	I'm already used to it	Used to express resignation.
80	见怪不怪	Nothing surprising	Used to normalize problems.
81	还能怎样	What else can we do	Used to express pessimism.
82	都一样	They're all the same	Used to express distrust toward alternatives.
83	别太认真	Don't take it too seriously	Used to lower discussion intensity.
84	看看就好	Just watch and move on	Used to encourage disengagement.
85	正常发挥	Normal performance	Used sarcastically to describe routine failure.
86	意料之中	As expected	Used to indicate predictable outcomes.
87	就这	That's it	Used to express disappointment.
88	这不很正常吗	Isn't this normal	Used sarcastically to normalize problems.
89	太阳底下没有新鲜事	Nothing new under the sun	Used to describe repetitive issues.
90	挡着我晒太阳了	Move aside you're blocking my sunlight	Used humorously to express detachment.
91	+1	+1	Used to express agreement.
92	同意楼上	Agree with the comment above	Used to endorse a previous comment.
93	复读机	Repeater machine	Used to describe repeated viewpoints.

No.	Original Expression (Chinese)	English Rendering	Description
94	复制粘贴	Copy and paste	Used to describe template dissemination.
95	模板来了	The template is here	Used to indicate repeated expression.
96	排队	Queue up	Used to signal collective participation.
97	接龙	Chain response	Used to describe sequential participation.
98	统一格式	Uniform format	Used to maintain expression consistency.
99	标准答案	The standard answer	Used sarcastically to describe consensus.
100	队形走起	Let the formation begin	Used to encourage collective repetition.
101	历史书写得最长的方法	Choosing the method that will fill the history books	Used to mock decisions with severe consequences.
102	君从叙拉古来	You come from Syracuse	Literary allusion used to satirize unrealistic political expectations.
103	天朝无敌	Our Celestial Empire is invincible	Used sarcastically to exaggerate nationalist narratives.
104	经济↓	economic ↓	Symbolic expression indicating economic decline while avoiding keywords.

Table 2: Original Expression (Chinese), English Rendering, and Description of 104 jokes

References

[1] Zúñiga, H., Molyneux, L., Zheng, P.: Social media, political expression, and political participation: Panel analysis of lagged and concurrent relationships. Journal of Communication **64**(4), 612–634 (2014) https://doi.org/10.1111/jcom.12103

[2] Papacharissi, Z.: Affective Publics: Sentiment, Technology, and Politics. Oxford Studies in Digital Politics. Oxford University Press, New York (2014). https://doi.org/10.1093/acprof:oso/9780199999736.001.0001

[3] Boyd, D.M., Ellison, N.B. Nicole B.: Social network sites: Definition, history, and scholarship. Journal of Computer-Mediated Communication **13**(1), 210–230 (2007) https://doi.org/10.1111/j.1083-6101.2007.00393.x

[4] Bennett, W.L., Segerberg, A.: The logic of connective action: Digital media

and the personalization of contentious politics. Information, Communication & Society **15**(5), 739–768 (2012) https://doi.org/10.1080/1369118X.2012.670661

[5] Pariser, E.: The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think. Penguin Books, USA (2012)

[6] Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., Starnini, M.: The echo chamber effect on social media. Proceedings of the National Academy of Sciences **118**(9), 2023301118 (2021) https://doi.org/10.1073/pnas.2023301118

[7] Bail, C.A., Argyle, L.P., Brown, T.W., Bumpus, J.P., Chen, H., Hunzaker, M.B.F., Lee, J., Mann, M., Merhout, F., Volfovsky, A.: Exposure to opposing views on social media can increase political polarization. Proceedings of the National Academy of Sciences **115**(37), 9216–9221 (2018) https://doi.org/10.1073/pnas.1804840115

[8] Roberts, S.T.: Behind the Screen: Content Moderation in the Shadows of Social Media. Yale University Press, ??? (2019)

[9] Kiene, C., Monroy-Hernández, A., Hill, B.M.: Surviving an "eternal september": How an online community managed a surge of newcomers. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16, pp. 1152–1156. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858356 . https://doi.org/10.1145/2858036.2858356

[10] Gorwa, R., Binns, R., Katzenbach, C.: Algorithmic content moderation: Technical and political challenges in the automation of platform governance. Big Data & Society **7**(1) (2020) https://doi.org/10.1177/2053951719897945

[11] Elmimouni, H., Rüller, S., Aal, K., Skop, Y., Abokhodair, N., Wulf, V., Tolmie, P.: Exploring algorithmic resistance: Responses to social media censorship in activism. Proc. ACM Hum.-Comput. Interact. **9**(2) (2025) https://doi.org/10.1145/3710970

[12] Elmimouni, H., Skop, Y., Abokhodair, N., Rüller, S., Aal, K., Weibert, A., Al-Dawood, A., Wulf, V., Tolmie, P.: Shielding or silencing?: An investigation into content moderation during the sheikh jarrah crisis. Proc. ACM Hum.-Comput. Interact. **8**(GROUP) (2024) https://doi.org/10.1145/3633071

[13] Ma, R., Kou, Y.: "how advertiser-friendly is my video?": Youtuber's socioeconomic interactions with algorithmic content moderation. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (2021) https://doi.org/10.1145/3479573

[14] Gillespie, T.: Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media. Yale University Press, ???

(2018). https://doi.org/10.12987/9780300235029

[15] Grimmelmann, J.: The virtues of moderation. Yale Journal of Law & Technology **17**, 42–109 (2015)

[16] DeVito, M.A., Birnholtz, J., Hancock, J.T., French, M., Liu, S.: How people form folk theories of social media feeds and what it means for how we study self-presentation. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems. CHI '18, pp. 1–12. Association for Computing Machinery, New York, NY, USA (2018). https://doi.org/10.1145/3173574.3173694 . https://doi.org/10.1145/3173574.3173694

[17] Ytre-Arne, B., Moe, H.: Folk theories of algorithms: Understanding digital irritation. Media, Culture & Society **43**(5), 807–824 (2021) https://doi.org/10.1177/0163443721992086

[18] Stoycheff, E.: Under surveillance: Examining facebook's spiral of silence effects in the wake of nsa internet monitoring. Journalism & Mass Communication Quarterly **93**(2), 296–311 (2016) https://doi.org/10.1177/1077699016630255

[19] Weeks, B.E., Halversen, A., Neubaum, G.: Too scared to share? fear of social sanctions for political expression on social media. Journal of Computer-Mediated Communication **29**(1), 041 (2024) https://doi.org/10.1093/jcmc/zmad041

[20] Daruwala, N.A.: Social media, expression, and online engagement: A psychological analysis of digital communication and the chilling effect in the uk. Frontiers in Communication **10**, 1565289 (2025) https://doi.org/10.3389/fcomm.2025.1565289

[21] Dal, A., Nisbet, E.C.: Walking through firewalls: Circumventing censorship of social media and online content in a networked authoritarian context. Social Media + Society **8**(4) (2022) https://doi.org/10.1177/20563051221141353

[22] Hiruncharoenvate, C., Lin, Z., Gilbert, E.: Algorithmically bypassing censorship on sina weibo with nondeterministic homophone substitutions. In: Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), vol. 9, pp. 150–158 (2015). https://doi.org/10.1609/icwsm.v9i1.14637

[23] Meng, B.: From steamed bun to grass mud horse: E gao as alternative political discourse on the chinese internet. Global Media and Communication **7**(1), 33–51 (2011) https://doi.org/10.1177/1742766510397938

[24] Pyle, C., Ellison, N.B., Andalibi, N.: Designing for discourse: Social media, socio-technical rhetorical strategies, and affirmative action discussions. In: Proceedings of the 2025 ACM Designing Interactive Systems Conference. DIS '25, pp. 1758–1773. Association for Computing Machinery, New York, NY, USA (2025). https://doi.org/10.1145/3715336.3735698 . https://doi.org/10.1145/3715336.3735698

[25] Halversen, A., Weeks, B.E.: Memeing politics: Understanding political meme creators, audiences, and consequences on social media. Social Media + Society **9**(4), 1–13 (2023) https://doi.org/10.1177/20563051231210570

[26] Ling, C., AbuHilal, I., Blackburn, J., De Cristofaro, E., Zannettou, S., Stringhini, G.: Dissecting the meme magic: Understanding indicators of virality in image memes. Proc. ACM Hum.-Comput. Interact. **5**(CSCW1) (2021) https://doi.org/10.1145/3449155

[27] King, G., Pan, J., Roberts, M.E.: How censorship in china allows government criticism but silences collective expression. American political science Review **107**(2), 326–343 (2013)

[28] Mina, A.X.: Batman, pandaman and the blind man: A case study in social change memes and internet censorship in china. Journal of Visual Culture **13**(3), 359–375 (2014)

[29] Al-Ani, B., Mark, G., Chung, J., Jones, J.: The egyptian blogosphere: a counter-narrative of the revolution. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work. CSCW '12, pp. 17–26. Association for Computing Machinery, New York, NY, USA (2012). https://doi.org/10.1145/2145204.2145213 . https://doi.org/10.1145/2145204.2145213

[30] Hansson, K., Pargman, T.C., Bardzell, S.: Materializing activism. Comput. Supported Coop. Work **30**(5–6), 617–626 (2021) https://doi.org/10.1007/s10606-021-09412-5

[31] Casula, P., Wong, R.Y.: "i see it, i scroll past it.": Exploring perceptions of social media political discourse among gen z young adult women in the u.s. Proc. ACM Hum.-Comput. Interact. **9**(2) (2025) https://doi.org/10.1145/3711007

[32] Gillani, N., Yuan, A., Saveski, M., Vosoughi, S., Roy, D.: Me, my echo chamber, and i: Introspection on social media polarization. In: Proceedings of the 2018 World Wide Web Conference. WWW '18, pp. 823–831. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3178876.3186130 . https://doi.org/10.1145/3178876.3186130

[33] Garimella, K., De Francisci Morales, G., Gionis, A., Mathioudakis, M.: Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In: Proceedings of the 2018 World Wide Web Conference. WWW '18, pp. 913–922. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE (2018). https://doi.org/10.1145/3178876.3186139 . https://doi.org/10.1145/3178876.3186139

[34] Grön, K., Nelimarkka, M.: Party politics, values and the design of social media services: Implications of political elites' values and ideologies to mitigating

of political polarisation through design. Proc. ACM Hum.-Comput. Interact. **4**(CSCW2) (2020) https://doi.org/10.1145/3415175

[35] Haimson, O.L., Delmonaco, D., Nie, P., Wegner, A.: Disproportionate removals and differing content moderation experiences for conservative, transgender, and black social media users: Marginalization and moderation gray areas. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (2021) https://doi.org/10.1145/3479610

[36] Ma, R., You, Y., Gui, X., Kou, Y.: How do users experience moderation?: A systematic literature review. Proc. ACM Hum.-Comput. Interact. **7**(CSCW2) (2023) https://doi.org/10.1145/3610069

[37] Jhaver, S., Appling, D.S., Gilbert, E., Bruckman, A.: "did you suspect the post would be removed?": Understanding user reactions to content removals on reddit. Proc. ACM Hum.-Comput. Interact. **3**(CSCW) (2019) https://doi.org/10.1145/3359294

[38] Feigenbaum, J., Jaggard, A.D., Wright, R.N.: Towards a formal model of accountability. In: Proceedings of the 2011 New Security Paradigms Workshop. NSPW '11, pp. 45–56. Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/2073276.2073282 . https://doi.org/10.1145/2073276.2073282

[39] Roberts, S.T.: Commercial content moderation: Digital laborers' dirty work (2016)

[40] Sullivan, J.: China's weibo: Is faster different? New Media & Society **16**, 24–37 (2014)

[41] Grinko, M., Qalandar, S., Randall, D., Wulf, V.: Nationalizing the internet to break a protest movement: Internet shutdown and counter-appropriation in iran of late 2019. Proc. ACM Hum.-Comput. Interact. **6**(CSCW2) (2022) https://doi.org/10.1145/3555205

[42] Feng, Y., Zhai, R., Sion, R., Carbunar, B.: A study of china's censorship and its evasion through the lens of online gaming. In: Proceedings of the 32nd USENIX Conference on Security Symposium. SEC '23. USENIX Association, USA (2023)

[43] Xiao, B.: Making the private public: Regulating content moderation under chinese law. Computer Law & Security Review **51**, 105893 (2023) https://doi.org/10.1016/j.clsr.2023.105893

[44] Wu, M., Sippe, J., Sivakumar, D., Burg, J., Anderson, P., Wang, X., Bock, K., Houmansadr, A., Levin, D., Wustrow, E.: How the great firewall of china detects and blocks fully encrypted traffic. In: 32nd USENIX Security Symposium (USENIX Security 23), pp. 2653–2670. USENIX Association, Anaheim, CA

(2023). https://www.usenix.org/conference/usenixsecurity23/presentation/wu-mingshi

[45] Kou, Y., Semaan, B., Nardi, B.: A Confucian Look at Internet Censorship in China. In: Bernhaupt, R., Dalvi, G., Joshi, A., Balkrishan, D.K., O'Neill, J., Winckler, M. (eds.) Lecture Notes in Computer Science. Human-Computer Interaction - INTERACT 2017, vol. LNCS-10513, pp. 377–398. Springer International Publishing, Bombay, India (2017). https://doi.org/10.1007/978-3-319-67744-6_25 . Part 6: Cultural Differences and Communication Technology. https://inria.hal.science/hal-01676171

[46] Lu, Z., Jiang, Y., Lu, C., Naaman, M., Wigdor, D.: The government's dividend: Complex perceptions of social media misinformation in china. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, pp. 1–12. Association for Computing Machinery, New York, NY, USA (2020). https://doi.org/10.1145/3313831.3376612 . https://doi.org/10.1145/3313831.3376612

[47] Perry, E.J.: Challenging the mandate of heaven: Popular protest in modern china. Critical Asian Studies **33**(2), 163–180 (2001)

[48] Perry, E.J.: Permanent rebellion? continuities and discontinuities in chinese protest. Popular Protest in China, 205–216 (2008)

[49] Whyte, M.: Myth of the Social Volcano: Perceptions of Inequality and Distributive Injustice in Contemporary China. Stanford University Press, ??? (2010)

[50] Bamman, D., O'Connor, B., Smith, N.: Censorship and deletion practices in chinese social media. First Monday **17**(3) (2012) https://doi.org/10.5210/fm.v17i3.3943

[51] Zhu, T., Phipps, D., Pridgen, A., Crandall, J.R., Wallach, D.S.: The velocity of censorship: high-fidelity detection of microblog post deletions. In: Proceedings of the 22nd USENIX Conference on Security. SEC'13, pp. 227–240. USENIX Association, USA (2013)

[52] Mou, Y., Atkin, D., Fu, H.: Predicting political discussion in a censored virtual environment. Political Communication **28**(3), 341–356 (2011) https://doi.org/10.1080/10584609.2011.572466

[53] Wang, Y., Mark, G.: Trust in online news: comparing social media and official media use by chinese citizens. In: Proceedings of the 2013 Conference on Computer Supported Cooperative Work. CSCW '13, pp. 599–610. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2441776.2441843 . https://doi.org/10.1145/2441776.2441843

[54] Kou, Y., Kow, Y.M., Gui, X.: Resisting the censorship infrastructure in china.

In: 50th Hawaii International Conference on System Sciences 2017 (HICSS-50), pp. 2332–2340 (2017). HICSS

[55] Wang, D., Mark, G.: Internet censorship in china: Examining user awareness and attitudes. ACM Trans. Comput.-Hum. Interact. **22**(6) (2015) https://doi.org/10.1145/2818997

[56] Jhaver, S., Bruckman, A., Gilbert, E.: Does transparency in moderation really matter? user behavior after content removal explanations on reddit. Proc. ACM Hum.-Comput. Interact. **3**(CSCW) (2019) https://doi.org/10.1145/3359252

[57] Cui, Y., Yamashita, N., Lee, Y.-C.: "we gather together we collaborate together": Exploring the challenges and strategies of chinese lesbian and bisexual women's online communities on weibo. Proc. ACM Hum.-Comput. Interact. **6**(CSCW2) (2022) https://doi.org/10.1145/3555148

[58] DeVos, A., Dhabalia, A., Shen, H., Holstein, K., Eslami, M.: Toward user-driven algorithm auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22. Association for Computing Machinery, New York, NY, USA (2022). https://doi.org/10.1145/3491102.3517441 . https://doi.org/10.1145/3491102.3517441

[59] Li, R., Kingsley, S., Fan, C., Sinha, P., Wai, N., Lee, J., Shen, H., Eslami, M., Hong, J.: Participation and division of labor in user-driven algorithm audits: How do everyday users work together to surface algorithmic harms? In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3582074 . https://doi.org/10.1145/3544548.3582074

[60] DeVito, M.A., Gergle, D., Birnholtz, J.: "algorithms ruin everything": #riptwitter, folk theories, and resistance to algorithmic change in social media. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems. CHI '17, pp. 3163–3174. Association for Computing Machinery, New York, NY, USA (2017). https://doi.org/10.1145/3025453.3025659 . https://doi.org/10.1145/3025453.3025659

[61] Rader, E., Gray, R.: Understanding user beliefs about algorithmic curation in the facebook news feed. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15, pp. 173–182. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2702123.2702174 . https://doi.org/10.1145/2702123.2702174

[62] Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., Sandvig, C.: "i always assumed that i wasn't really that

close to [her]": Reasoning about invisible algorithms in news feeds. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15, pp. 153–162. Association for Computing Machinery, New York, NY, USA (2015). https://doi.org/10.1145/2702123.2702556 . https://doi.org/10.1145/2702123.2702556

[63] Kingsley, S., Sinha, P., Wang, C., Eslami, M., Hong, J.I.: "give everybody [..] a little bit more equity": Content creator perspectives and responses to the algorithmic demonetization of content associated with disadvantaged groups. Proc. ACM Hum.-Comput. Interact. **6**(CSCW2) (2022) https://doi.org/10.1145/3555149

[64] Ma, R., Kou, Y.: "i'm not sure what difference is between their content and mine, other than the person itself": A study of fairness perception of content moderation on youtube. Proc. ACM Hum.-Comput. Interact. **6**(CSCW2) (2022) https://doi.org/10.1145/3555150

[65] Delmonaco, D., Mayworm, S., Thach, H., Guberman, J., Augusta, A., Haimson, O.L.: "what are you doing, tiktok?": How marginalized social media users perceive, theorize, and "prove" shadowbanning. Proc. ACM Hum.-Comput. Interact. **8**(CSCW1) (2024) https://doi.org/10.1145/3637431

[66] Shklovski, I., Kotamraju, N.: Online contribution practices in countries that engage in internet blocking and censorship. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '11, pp. 1109–1118. Association for Computing Machinery, New York, NY, USA (2011). https://doi.org/10.1145/1978942.1979108 . https://doi.org/10.1145/1978942.1979108

[67] DeVito, M.A.: Adaptive folk theorization as a path to algorithmic literacy on changing platforms. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (2021) https://doi.org/10.1145/3476080

[68] Vaccaro, K., Sandvig, C., Karahalios, K.: "at the end of the day facebook does what itwants": How users experience contesting algorithmic content moderation. Proc. ACM Hum.-Comput. Interact. **4**(CSCW2) (2020) https://doi.org/10.1145/3415238

[69] Vaccaro, K., Xiao, Z., Hamilton, K., Karahalios, K.: Contestability for content moderation. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (2021) https://doi.org/10.1145/3476059

[70] Atreja, S., Im, J., Resnick, P., Hemphill, L.: Appealmod: Inducing friction to reduce moderator workload of handling user appeals. Proc. ACM Hum.-Comput. Interact. **8**(CSCW1) (2024) https://doi.org/10.1145/3637296

[71] Eslami, M., Karahalios, K., Sandvig, C., Vaccaro, K., Rickman, A., Hamilton, K., Kirlik, A.: First i "like" it, then i hide it: Folk theories of social

feeds. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16, pp. 2371–2382. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2858036.2858494 . https://doi.org/10.1145/2858036.2858494

[72] Choi, Y., Kang, E.J., Lee, M.K., Kim, J.: Creator-friendly algorithms: Behaviors, challenges, and design opportunities in algorithmic platforms. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems. CHI '23. Association for Computing Machinery, New York, NY, USA (2023). https://doi.org/10.1145/3544548.3581386 . https://doi.org/10.1145/3544548.3581386

[73] Karizat, N., Delmonaco, D., Eslami, M., Andalibi, N.: Algorithmic folk theories and identity: How tiktok users co-produce knowledge of identity and engage in algorithmic resistance. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (2021) https://doi.org/10.1145/3476046

[74] Sleeper, M., Balebako, R., Das, S., McConahy, A.L., Wiese, J., Cranor, L.F.: The post that wasn't: exploring self-censorship on facebook. CSCW '13, pp. 793–802. Association for Computing Machinery, New York, NY, USA (2013). https://doi.org/10.1145/2441776.2441865 . https://doi.org/10.1145/2441776.2441865

[75] Das, S., Kramer, A.D.I.: Self-censorship on facebook. Proceedings of the International AAAI Conference on Web and Social Media (2013)

[76] Vitak, J., Kim, J.: "you can't block people offline": examining how facebook's affordances shape the disclosure process. In: Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing. CSCW '14, pp. 461–474. Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2531602.2531672 . https://doi.org/10.1145/2531602.2531672

[77] Hakami, Z., Feng, Y., Carbunar, B.: Cooperative dynamics of censorship, misinformation, and influence operations: Insights from the global south and u.s. Proc. ACM Hum.-Comput. Interact. **9**(7) (2025) https://doi.org/10.1145/3757572

[78] Chancellor, S., Pater, J.A., Clear, T., Gilbert, E., De Choudhury, M.: #thyghgapp: Instagram content moderation and lexical variation in pro-eating disorder communities. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW '16, pp. 1201–1213. Association for Computing Machinery, New York, NY, USA (2016). https://doi.org/10.1145/2818048.2819963 . https://doi.org/10.1145/2818048.2819963

[79] Moran, R.E., Grasso, I., Koltai, K.: Folk theories of avoiding content moderation: How vaccine-opposed influencers amplify vaccine opposition on instagram. Social Media+ Society **8**(4), 20563051221144252 (2022)

[80] Ye, W., Zhao, L.: "i know it's sensitive": Internet censorship, recoding, and the

sensitive word culture in china. Discourse, Context & Media **51**, 100666 (2023) https://doi.org/10.1016/j.dcm.2022.100666

[81] Lee, S.-y.: Surviving online censorship in china: Three satirical tactics and their impact. The China Quarterly **228**, 1061–1080 (2016)

[82] Luqiu, L.R.: The cost of humour: Political satire on social media and censorship in china. Global Media and Communication **13**(2), 123–138 (2017)

[83] Mayworm, S., DeVito, M.A., Delmonaco, D., Thach, H., Haimson, O.L.: Content moderation folk theories and perceptions of platform spirit among marginalized social media users. Trans. Soc. Comput. **7**(1–4) (2024) https://doi.org/10.1145/3632741

[84] Boyd, D., Marwick, A.E.: Social privacy in networked publics: Teens' attitudes, practices, and strategies. In: A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society (2011)

[85] Marwick, A.E., Boyd, D.: Networked privacy: How teenagers negotiate context in social media. New media & society **16**(7), 1051–1067 (2014)

[86] Petronio, S., Child, J.T., Hall, R.D.: Communication privacy management theory: Significance for interpersonal communication. In: Engaging Theories in Interpersonal Communication, pp. 314–327. Routledge, ??? (2021)

[87] Clark, H.H.: Using Language. Cambridge university press, ??? (1996)

[88] Clark, H.H., Brown, K.: Context and common ground. Concise encyclopedia of philosophy of language and linguistics, 85–87 (2006)

[89] Marwick, A.E., Boyd, D.: I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. New media & society **13**(1), 114–133 (2011)

[90] Andrejevic, M.: The work of watching one another: Lateral surveillance, risk, and governance. Surveillance & Society **2**(4) (2004)

[91] Cotter, K.: Playing the visibility game: How digital influencers and algorithms negotiate influence on instagram. New Media & Society **21**(4), 895–913 (2019) https://doi.org/10.1177/1461444818815684 https://doi.org/10.1177/1461444818815684

[92] Marwick, A.E., boyd: I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. New Media & Society **13**(1), 114–133 (2011) https://doi.org/10.1177/1461444810365313 https://doi.org/10.1177/1461444810365313

[93] Litt, E., Hargittai, E.: The imagined audience on social network sites. Social

Media+ Society **2**(1), 2056305116633482 (2016)

[94] Dresner, E., Herring, S.C.: Functions of the nonverbal in cmc: Emoticons and illocutionary force. Communication theory **20**(3), 249–268 (2010)

[95] Li, Z., Nardi, B.: "there should be more than one voice in a healthy society": Infrastructural violence and totalitarian computing in china. Proc. ACM Hum.-Comput. Interact. **5**(CSCW2) (2021) https://doi.org/10.1145/3476070

[96] Bucher, T.: The algorithmic imaginary: exploring the ordinary affects of facebook algorithms. Information, Communication & Society **20**(1), 30–44 (2017) https://doi.org/10.1080/1369118X.2016.1154086 https://doi.org/10.1080/1369118X.2016.1154086

[97] Vitak, J.: The impact of context collapse and privacy on social network site disclosures. Journal of Broadcasting & Electronic Media **56**(4), 451–470 (2012) https://doi.org/10.1080/08838151.2012.732140

[98] Klonick, K.: The new governors: The people, rules, and processes governing online speech. Harv. L. Rev. **131**, 1598 (2017)

[99] Kiesler, S., Siegel, J., McGuire, T.W.: Social psychological aspects of computer-mediated communication. American psychologist **39**(10), 1123 (1984)

[100] Yang, T.Z.: Normalization of censorship: Evidence from china. The Journal of Politics **87**(4), 1227–1242 (2025)

[101] Draper, N.A., Turow, J.: The corporate cultivation of digital resignation. New Media & Society **21**(8), 1824–1839 (2019) https://doi.org/10.1177/1461444819833331 https://doi.org/10.1177/1461444819833331